

ESTIMATION OF GENERALIZED SIMPLE MEASUREMENT ERROR
MODELS WITH INSTRUMENTAL VARIABLES

By

JEFFREY RAY THOMPSON

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2003

Copyright 2003

by

Jeffrey Ray Thompson

I dedicate this work to my parents, Ron and Ellie, and my sister, Jennifer.

ACKNOWLEDGMENTS

First and foremost, I would like to thank Dr. Randy Carter for all the help required in writing this dissertation. Dr. Carter has been an outstanding advisor, colleague, and friend and his patience and willingness to teach are unparalleled. Without his guidance, this dissertation would not have been possible. He has given me knowledge, both pertaining to statistics and otherwise, that I will be able to apply throughout my career and many other facets of my life, as well. Further I would like to thank the members of my committee: Dr. Malay Ghosh, Dr. Ramon Littell, Dr. Ron Randles, and Dr. Mike Resnick. Either as teachers in my courses or mentors outside of the classroom, each member of my committee has played a unique and special role in my time here as a graduate student.

Second, I would like to thank my family. My parents, Ellie and Ron, and sister, Jennifer (and I guess her husband, Matt, too), have provided an infinite amount of support and encouragement throughout this endeavor. Even at my most frustrating times, they all had faith in me and gave me the will to succeed. There are no words to describe how thankful I am for such unwavering support. I guess I should just say, "I love you all very much!"

I would like to thank all my friends from school, especially all my officemates over the years and those who I repeatedly went to with my questions. Whether it was studying for the comprehensive or qualifying exams together, helping me through the toughest of classes, or simply having someone who has been there before to relate to, I know I would not have made it this far without them.

Further I would like to thank those friends of mine outside the world of statistics. Both my life-long friends and the friends I have made here since being in

Gainesville have been crucial to my success as a graduate student for being there with me when I needed to blow off steam or helping me put life into perspective when things with my studies got difficult.

I would also like to thank those who played early roles in placing me on this career path. Whether it was taking me under their wing or showing me what a great teacher really is, to all those early influences, I owe a great amount of gratitude.

Finally, I would like to thank my cat Bobbi, who unconditionally was always there to greet me at my door at the end of the most trying days and also could help put a smile on my face at the times I needed it most.

TABLE OF CONTENTS

	<u>page</u>
ACKNOWLEDGMENTS	iv
ABSTRACT	viii
CHAPTER	
1 INTRODUCTION: MEASUREMENT ERROR MODELS	1
1.1 Common Occurrences of the Measurement Error Problem	1
1.2 Motivating Examples	2
1.3 A Known Solution	6
2 LITERATURE REVIEW AND BACKGROUND	8
2.1 The Multivariate Measurement Error Model	8
2.2 The Simple Linear ME Model with Normal Errors and True Values	9
2.2.1 Non-Identifiability and Asymptotic Bias of the OLS Estimator	11
2.2.2 Additional Information Required for Identifiability	13
2.3 The Multivariable Linear ME Model Under Normality	22
2.3.1 Non-Identifiability and Bias	23
2.3.2 Additional Information Required	24
2.4 Nonlinear ME Models	27
2.4.1 Normal Theory Models: Non-Identifiability and Bias	28
2.4.2 Normal Theory Models: Additional Information Required	29
2.4.3 Exponential Family Models: Non-Identifiability and Bias	32
2.4.4 Exponential Family Models: Addition Information Required	34
3 GENERALIZED SIMPLE M.E. MODEL WITH INSTRUMENTAL VARIABLES	42
3.1 The Model	42
3.2 Applications of the GSME Model	47
3.2.1 Exponential Family General Linear Models	48
3.2.2 Multinomial Model	49
3.2.3 Survival Time	51
3.2.4 Multivariate Regression Model	53
3.2.5 Nonlinear Model	55
3.3 Model Identification	59

4	ESTIMATION OF THE G.S.M.E. MODEL	66
4.1	The Univariate Case	69
4.1.1	Translation to Latent Class Model	69
4.1.2	Estimation of Model Parameters	74
4.1.3	Integration Examples of the Categorization Step	84
4.2	Generalization to Estimating a Multivariate Model	90
4.3	Asymptotic Properties	94
4.4	Consistent Initial Estimates	110
5	APPLICATION TO MOTIVATING EXAMPLE	112
6	OTHER POSSIBLE APPLICATIONS	121
6.1	Application for MCHERDC with Developmental Delay and Dis- ability Outcome	121
6.2	Application to Alzheimer's Disease	126
7	DISCUSSION AND FUTURE WORK	130
7.1	Discussion of the GSME Model	130
7.2	Monte Carlo Study and Asymptotic Efficiency	132
7.3	The GSME Model with Error-Free Covariates	133
7.4	An Alternate Method	136
	REFERENCES	143
	BIOGRAPHICAL SKETCH	146

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

ESTIMATION OF GENERALIZED SIMPLE MEASUREMENT ERROR
MODELS WITH INSTRUMENTAL VARIABLES

By

Jeffrey Ray Thompson

August 2003

Chair: Randy L. Carter

Major Department: Statistics

Measurement error (ME) models are used in situations where at least one independent variable in the model is imprecisely measured. Having at least one independent variable measured with error leads to an unidentified model and a bias in the naive estimate of the effect of the variable that is measured with error. One way to correct these problems is through the use of an instrumental variable (IV). An IV is one that is correlated with the unknown, or latent, true variable, but uncorrelated with the measurement error of the unknown truth and the model error. An IV provides the identifying information in our method of estimating the parameters for generalized simple measurement error (GSME) models. The GSME model is developed and it is shown how many well studied ME models with one predictor can fit into its framework. Included in these are linear, generalized linear, nonlinear, multinomial, multivariate regression, and other ME models. The GSME model, by design, can handle situation for continuous, discrete, and categorical observable, or manifest, variables. We provide theorems that give conditions under which the GSME model is identified. The initial step in our estimation method is to “categorize” all continuous and discrete variables. Categorical variables remain

unchanged. Assuming conditional independence given the latent variable, the joint distribution of the categorized manifest variables and any that were already categorical is the product of the conditional cell probabilities and conditional distributions of the categorized continuous and discrete manifest variables summing over the categorical values of the latent variable. Maximum likelihood estimates of the joint categorical distribution are used to solve nonlinear equations for the parameters of interest which enter through the conditional probabilities. Estimated generalized nonlinear least squares is used to solve the equations for the parameters of interest. We show that our estimators have favorable asymptotic properties and develop methods of inference for them. We show how many commonly studied ME model problems fit into the general framework developed and how they can be solved using our method.

CHAPTER 1 INTRODUCTION: MEASUREMENT ERROR MODELS

1.1 Common Occurrences of the Measurement Error Problem

Measurement error models are models with at least one independent variable that is measured with error. Aside from the term *measurement error models*, this area of statistics is known by several other names: *errors in variables* and *regression with errors in x* are but two others [56]. The measurement error problem occurs often in practice. The medical field, agriculture, and econometrics are just a few areas that contain problems involving predictor variables which are contaminated by measurement error.

The general measurement error problem occurs often in medical research when the association between an imperfectly measured predictor variable and a response variable is of interest. One of many examples involves the occurrence of Alzheimer's disease. Alzheimer's disease affects roughly fifteen million people worldwide and over four million in the United States alone [32]. One in ten persons over the age of sixty-five has the disease and it also is the primary cause of dementia in the elderly [4]. It is thought that the level of aluminum deposits, which may build up in the brain over time, have an effect on the likelihood of an individual developing Alzheimer's [12]. If one were to estimate the association between the likelihood of Alzheimer's Disease and the level of aluminum deposits in the brain, problems would arise because a perfectly accurate measure of aluminum levels is simply not attainable. Measurement error models have also been used to assess the reliability and validity of a measuring method in the absence of a true "gold standard" method. Carter [17], for example, assessed the reliability and

validity of a measure of specific activity of the enzyme sucrase that was obtained from a homogenated sample of the small intestine of intestinal bypass patients.

There are also numerous examples of measurement error models occurring in agriculture. In one such example, considered by Fuller [24], the experience and education of a farmer were used as predictors in estimating farm size. In this example, random error was added to each variable in order to protect the confidentiality of the respondents. Therefore, two types of errors existed in the explanatory variables: the error added to protect confidentiality and that of the original responses of experience and education as stated by individual farmers.

In economics, one might be interested in expressing, say, the net exports of a commodity as a function of the available resources to a country. Klepper and Leamer [36] faced this problem when estimating the amount of machinery exported by a country as a linear function of the country's land, labor, and capital. They could not apply usual linear regression techniques to correctly estimate the amount of machinery exported, because these three explanatory variables were, in their words, "doubtlessly measured with error" [36, p. 180].

1.2 Motivating Examples

The Radiations Effects Research Foundation (RERF) in Hiroshima, Japan, is involved in radiation dose-response studies using the Atomic-bomb survivor data. Information from residents of both Hiroshima and Nagasaki was obtained from mostly direct interviews with survivors, but also with interviews from family members and acquaintances. Shielding information and distance from the bombs hypocenter were obtained from such interviews and were used to estimate each individual's radiation dose [41]. These estimates of individual radiation doses have substantial random errors [43]. The random errors in dose estimates are due largely to uncertainties regarding the survivors' location from the hypocenter and shielding [42] and imperfect estimation of true dose from these variables and other error-free

variables. Sposto et al. [54] estimated and discussed the magnitude of the errors in the radiation dose estimates.

Pierce et al. [43] was interested, primarily, in cancer incidence and mortality rates, where the number of cases in a given time interval was assumed to follow a Poisson distribution, but they also gave some consideration to the analysis of data for chromosomal aberrations, in which the number of cases was assumed to be binomial. For example, in the analysis of chromosomal aberration data, the outcome was the proportion of about 100 examined cells that exhibit an aberration, and this outcome was assumed to follow a binomial distribution [42, p. 279]. Modeling these outcomes involved mainly generalized linear models with varying assumptions depending on the outcome of interest, but they also briefly mention applications in survival analysis, using a proportional hazards model. One of the primary concerns was the incidence of leukemia which was modeled as the hazard function of a grouped survival analysis. The outcome was time until the diagnosis of leukemia and it will serve as the primary outcome of interest in our discussion. Pierce et al. [43] discussed examples where the number of cases follows a binomial (for chromosomal aberration data) and a Poisson (for cancer incidence and mortality) distribution, but in both cases, they assumed an identity link function so the Y 's, i.e., the outcomes, have conditional expectations that are linear in x . A Weibull distribution, with slightly different parameters depending on city, was assumed for true radiation dose, x , and multiplicative errors were assumed for estimated dose that were independent from individual to individual. The unobservable x was the true radiation dose and the estimated value, X , was the estimated dose from the Dosimetry System 1986, henceforth, DS86. Their approach involved replacing each individuals' true dose, x_i , by an estimate of $E(x_i | X_i)$ and carrying out a weighted least squares analysis as if the estimate was the true dose. Estimation of $E(x_i | X_i)$ was possible because the distribution of X_i

given x_i was assumed to follow a lognormal distribution with a known coefficient of variation. (The authors recognized that the coefficient of variation actually is not known, and performed sensitivity analyses to assess the potential impact of a violation of this assumption [43, p. 352].) A value for $E(x | X)$ was obtained from the relationship $f(x | X) \propto f(x)f(X | x)$. A lognormal distribution with known coefficient of variation means that $\log(X)$ is assumed to be normally distributed with conditional mean $\log(x)$ having a coefficient of variation, assumed in this case, of 30% and 40%. Under this range of values for a lognormal distribution, the standard deviation of $\log(X)$ is approximately equal to the coefficient of variation. For the precise relationship for lognormal models, see Pierce et al. [42, p. 278]. This approach is an example use of a method known as the regression calibration method, which will be discussed briefly in the next section and more thoroughly in the next chapter. The authors also gave results for two distributions which were not lognormal (namely, contaminated lognormal and normal). Ultimately, their recommendation was to use a lognormal distribution with a 35% coefficient of variation.

There are two key extensions to their method that they mention but do not apply. The first concerned the need to add covariates measured without error to their model. In the current context, these covariates would be sex, city, and age ATB and are denoted collectively by Z . Pierce et al. [43] mentioned that in the presence of such variables, the estimate of $E(x | X)$ should be replaced in their method with an estimate of $E(x | X, Z)$. If Z is in the model for $E(Y)$, they state that using $E(x | X, Z)$ should reduce the bias in parameter estimates. They do not pursue this idea further.

Another useful suggestion they made, but did not pursue, was the use of a sign of acute radiation effects as a possible instrumental variable (IV) in order to avoid the need to assume a known coefficient of variation. One possible sign of acute

radiation effects is the presence or absence of severe epilation. Severe epilation is defined to be scalp hair loss of more than 67% [41]. This variable does seem to satisfy the requirements of an IV. Severe epilation is related to the true radiation dose but assumably independent of the measurement error in dose and of the model error and measurement error in Y , if any. Not only do Pierce et al. not pursue the use of an IV, they do not make any suggestions as to how to incorporate an IV into their method.

Another possible application of our methods, comes from the area of maternal and child health epidemiology. Consider as an outcome, an indicator of early childhood disability or developmental delay (DDD). It is thought that developmental delay and disability in early childhood may be related to gestational age (GA) (see, for example, [66, 30, 7]). Because GA is known to be measured with error, however, researchers typically use birth weight as the predictor of interest (see, for example, [60, 61]). Such studies are useful and informative in their own right, but do not answer questions concerning the relationship between the likelihood of DDD and prematurity. Thus, the use of GA would be an ideal variable to assess this relationship, except, in most existing studies, the fact GA is often so inaccurately estimated is not accounted for.

The methods presented here may offer a solution to this dilemma. Since there exist multiple methods for estimating GA (see, for example, [53, 39]), a possible consideration for an IV may be a second measure of GA, obtained from an independent method. In Chapter 6 we discuss this possible application and show how the outcome of interest, DDD, may be applied to our methods in a future study that would estimate the effects of prematurity on a child's risk of DDD in the first three years of life.

1.3 A Known Solution

The solution provided by Pierce et al. [43] in the atomic bomb study is an example application of a methodology first proposed by Prentice [44] in the context of a proportional hazards model and subsequently generalized by Carroll et al. [13]. The method is called the regression calibration method. It amounts to estimating $E(x | \cdot)$ and then substituting the estimates for values of x in the analysis that would have been done if x had been observed without error. Unfortunately, $E(x | \cdot)$ cannot be estimated without additional information. The additional information needed can exist in the form of a known or independently estimated parameter (Pierce et al. [43], for example assumed that the coefficient of variation of the conditional distribution of x given X was known), repeated measures of x , or an instrumental variable (i.e., a variable that is related to x but independent of the measurement error in x and independent of the model error and measurement error in Y , if any). The regression calibration method can be applied to either linear or nonlinear models.

It should be noted that $E(x | \cdot) = E(x | X)$ if there are no covariates in the model that are measured without error (Z) and no instrumental variables (W , say). If there are error free covariates but no IV then $E(x | \cdot) = E(x | X, Z)$, $E(x | \cdot) = E(x | X, W)$ if there is an IV but no error free covariates, and $E(x | \cdot) = E(x | X, Z, W)$ in the presence of both.

The methods proposed by Carroll et al. [13] for IV estimation rely heavily on two, rather restrictive, assumptions. The first is that when their methods are applied to generalized linear models, they consider only canonical link functions. Therefore, they write $E(Y | Z, x) = b'(\beta_{Y|1Zx} + Z'\beta_{Y|1Zx} + x'\beta_{Y|1Zx})$, where $\beta_{Y|1Zx}$ is the coefficient of 1, i.e., the intercept, in the regression of Y on 1, Z , and x , $\beta_{Y|1Zx}$ is the coefficient of Z' in the same regression, and $\beta_{Y|1Zx}$ is the coefficient of x' . The other restriction is that there is a linear regression of x on (Z, W, X) . In other

words, the assumption that the covariates measured with error and any covariates measured without error, the instrumental variables, and the observed measurements of the x 's have the following relation:

$$E(x \mid Z, W, X) = \beta_{x|1ZW}X + Z'\beta_{x|1ZW}X + W'\beta_{x|1ZW}X + X'\beta_{x|1ZW}X.$$

The coefficients are defined similarly as those above. This assumption of a linear relation between these variables is usually incorrect if x are categorical. This is easily seen when x is dichotomous. Also, they assume that $X = x + u$ and $E(x - X \mid Z, x, W) = 0$, and, therefore, that X is an unbiased measurement of x . The methods suggested in this proposal shall require no such assumptions of a canonical link (when applied to GLM's), a strictly linear relation between x , the observed covariates, and IV's, nor that of an unbiased measurement of x . The key assumption is that of conditional independence, which is also an assumption made by Carroll et al. [13].

In Chapter 2, we define the general, multivariate measurement error model for the mean function of a response variable. After which, we will look at several estimators under specific measurement error models, starting with the simplest case, and discuss their properties.

CHAPTER 2 LITERATURE REVIEW AND BACKGROUND

2.1 The Multivariate Measurement Error Model

Let $\{\underline{y}_i\}_{i=1}^{\infty}$ and $\{\underline{e}_i\}_{i=1}^{\infty}$ be sequences of p -dimensional random column vectors and let $\{\underline{x}_i\}_{i=1}^{\infty}$ be a sequence of k -dimensional independent column vectors. Further, let $\underline{\beta} \in \psi \subset R^{kp}$ be a $kp \times 1$ column vector of parameters and $\underline{f}$ be a p -dimensional vector whose components are real valued Borel measurable functions mapping $R^k \times \psi$ into R^1 and describing $E(Y_i | \underline{x}_i)$ as a function of $\underline{x}_i$. Then

$$\underline{Y}_i = \underline{f}(\underline{x}_i; \underline{\beta}) + \underline{e}_i, \quad i = 1, 2, \dots, \quad (2.1)$$

where $\underline{e}_i$ may be a combination of model error and measurement error, if it exists. Also, assume

$$E(e_{ij}) = E(x_{il}e_{i'j}) = 0 \quad (2.2)$$

for all $j = 1, 2, \dots, p$, $l = 1, 2, \dots, k$, $i = 1, 2, \dots$, and $i' = 1, 2, \dots$. If the elements of $\underline{\beta}$ are not functionally related and

$$\underline{f}(\underline{x}_i; \underline{\beta}) = \underline{B}\underline{x}_i, \quad i = 1, 2, \dots, \quad (2.3)$$

where $\underline{B}$ is a $p \times k$ matrix of parameters formed by placing the first k elements of the $kp \times 1$ vector $\underline{\beta}$ in the first row of $\underline{B}$, the next k in the second row, and so on, then the model defined by Equations 2.1 and 2.2 is called the *multivariate nonlinear regression model*. With the addition of Equation 2.3, the model is called the *multivariate linear regression model*. Note that if the model contains an intercept, the first element of $\underline{x}_i$ is unity. It is well known that under the assumption of normally distributed model error, the least squares estimator for $\underline{B}$

(note: least squares does not require normality assumption) when the $\underline{x}_i$ are fixed and the maximum likelihood estimator when the $\underline{x}_i$ are random drawings from a normal distribution are equal and both unbiased estimates [24]. It sometimes is the case that the vectors $\underline{x}_i$, $i = 1, 2, \dots$, in Equation 2.3 are not directly observable, but are observable with error. This leads to the so called linear errors in variables model.

When the vectors $\underline{x}_i$, $i = 1, 2, \dots$, are not directly observable, one may observe, in a finite sample,

$$\underline{X}_i = \underline{x}_i + \underline{u}_i, \quad i = 1, 2, \dots, n. \quad (2.4)$$

In addition to the first assumption in 2.2, we assume that the k-vectors $\underline{u}_i$, $i = 1, 2, \dots, n$, of measurement error random variables are such that

$$E(u_{il}) = E(u_{il}x_{il'}) = E(u_{il}y_{il'}) = 0 \quad (2.5)$$

for all $j = 1, 2, \dots, p$, $l = 1, 2, \dots, k$, $l' = 1, 2, \dots, k$, $i = 1, 2, \dots, n$, and $i' = 1, 2, \dots, n$. The model defined by Equations 2.1, 2.2, 2.4, and 2.5 is called the *general measurement error* (henceforth, ME) *model*. With the addition of Equation 2.3, the model is called the *linear measurement error model* (linear ME). When the unobservable $\underline{x}_i$ are fixed, the model is known as the *functional model* and when they are nonconstant random vectors, the model is known as the *structural model*.

2.2 The Simple Linear ME Model with Normal Errors and True Values

The majority of this section will be devoted to the structural model, but it is none the less important to distinguish between the structural and functional models. Again, the $(\underline{x}_1, \underline{x}_2, \dots)$ in Equation 2.1 may be observations on non-random variables in which case the model is defined to be the functional model. In this case, the $\underline{x}_i$ can be thought of as a constant vector and they appear as parameters in the distribution function of $(\underline{Y}_i, \underline{X}_i)$. In the structural model with normal true values, the $\underline{x}_i$ are assumed to be independent drawings from a $N(\underline{\mu}_x, \Sigma_{xx})$

distribution. The name “structural” comes from the fact that these models describe the structure of the specified relationship between the random variables $\underline{x}$ and $\underline{Y}$.

For illustrative purposes, we shall begin with the simple linear ME model which contains one dependent and one independent variable and has normally distributed model error. Consider the model defined by

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad (2.6)$$

$$X_i = x_i + u_i, \quad (2.7)$$

where $i = 1, 2, \dots, n$, e_i are independent $N(0, \sigma_{ee})$ and may potentially be a combination of model and measurement error, and u_i is a $N(0, \sigma_{uu})$ random variable. So, at this point, we have the following assumptions:

$$\begin{bmatrix} x_i \\ e_i \\ u_i \end{bmatrix} \sim^{iid} N \left(\begin{bmatrix} \mu_x \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & \sigma_{xe} & \sigma_{xu} \\ \sigma_{xe} & \sigma_{ee} & \sigma_{eu} \\ \sigma_{xu} & \sigma_{eu} & \sigma_{uu} \end{bmatrix} \right), \quad (2.8)$$

where $\sim^{iid}$ means “independently and identically distributed” and N means “normally” distributed. For some estimation purposes, it may be assumed that x_i , e_i , and u_i are independent for all i . When there is no measurement error (i.e., the classical regression model), it is well known that the least squares estimator of β_1 is

$$\hat{\beta}_1 = \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{-1} \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}).$$

This estimator is also the maximum likelihood estimator and is unbiased for β_1 in both the functional and structural models.

In addition to the assumptions in (2.8) that the $(x_1, x_2, \dots, x_n)$ in Equations 2.6 and 2.7 are random (i.e., they are drawn from a $N(\mu_x, \sigma_{xx})$ distribution) we may in some applications also assume that x_i , e_i , and u_i are independent for all i .

Then we have that the vector

$$(x_i, e_i, u_i)' \sim^{iid} N[(\mu_x, 0, 0)', \text{diag}(\sigma_{xx}, \sigma_{ee}, \sigma_{uu})], \quad (2.9)$$

where $\text{diag}()$ represents a diagonal matrix with the given elements on the diagonal. Then the vector $(Y_i, X_i)'$, where Y_i is defined by Equation 2.6 and X_i is defined by Equation 2.7, has a bivariate normal distribution with mean vector $E\{(Y_i, X_i)'\} = (\mu_Y, \mu_X)' = (\beta_0 + \beta_1\mu_x, \mu_x)'$ and variance-covariance matrix

$$\text{Var}\{(Y_i, X_i)'\} = \begin{bmatrix} \sigma_{YY} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{XX} \end{bmatrix} = \begin{bmatrix} \beta_1^2\sigma_{xx} + \sigma_{ee} & \beta_1\sigma_{xx} \\ \beta_1\sigma_{xx} & \sigma_{xx} + \sigma_{uu} \end{bmatrix}.$$

2.2.1 Non-Identifiability and Asymptotic Bias of the OLS Estimator

Under the model defined by Equations 2.6 and 2.7 and under the assumptions stated in (2.9), using the observed variables, one might naively estimate the regression coefficient, β_1 , as

$$\hat{\beta}_{1,OLS} = \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right]^{-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}). \quad (2.10)$$

By the properties of the bivariate normal distribution, this naive estimator has expected value $E\{\hat{\beta}_{1,OLS}\} = \sigma_{XX}^{-1}\sigma_{XY} = \beta_1(\sigma_{xx} + \sigma_{uu})^{-1}\sigma_{xx}$. One can see that, because the denominator is inflated by σ_{uu} , the least squares regression coefficient is biased towards zero, and that the bias does not vanish with increasing sample size. It is important to note that key assumptions used to derive the properties of this estimator were that the measurement error, u_i , was independent of both the true values, x_i , and the model errors, e_i . This bias towards zero in the regression coefficient is also referred to as attenuation of the coefficient toward zero. The ratio $\kappa_{xx} = \sigma_{XX}^{-1}\sigma_{xx}$, which defines the degree of attenuation, is known as the reliability ratio, and measures the reliability of X as a measurement of x . Note that the

ratio ranges from zero to 1.0 with larger values indicating greater reliability of measurement.

Another important concept that we shall discuss is that of identifiability of a model. Fuller [24, p. 9-10] provides the following definition of *identifiability*: “Let $\mathbf{Z}$ be the vector of observable random variables and let $F_{\mathbf{Z}}(\mathbf{a} : \theta)$ be the distribution function of $\mathbf{Z}$, evaluated at $\mathbf{a}$, for the given parameter θ in the parameter space Θ . The parameter θ is identified if, for any $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$, $\theta_1 \neq \theta_2$ implies that $F_{\mathbf{Z}}(\mathbf{a} : \theta_1) \neq F_{\mathbf{Z}}(\mathbf{a} : \theta_2)$ for some $\mathbf{a}$.” Further, we say that if the vector θ is identified, then the “model” is identified [24, p. 10]. By the term “model,” we mean a specification of the variables and parameters of interest, the relationships among the variables, and the assumptions about the stochastic properties of the random variables. For the structural model defined by Equations 2.6 and 2.7, the vector of unknown parameters, θ , is given by $(\mu_x, \sigma_{xx}, \sigma_{ee}, \sigma_{uu}, \beta_0, \beta_1)$. Under the assumptions of the stated model, the observations (Y_i, X_i) have a bivariate normal distribution. Properties of the bivariate normal distribution state that it shall be completely characterized by the elements of its mean vector and variance-covariance matrix. Thus the distribution of (Y_i, X_i) is characterized by the five parameters in its mean vector and variance-covariance matrix, namely, $(\mu_Y, \mu_X, \sigma_{YY}, \sigma_{XX}, \sigma_{XY})$. Because the model contains six different parameters in the vector θ , it is not possible to find a unique relationship between θ and the parameter vector of the distribution of (Y_i, X_i) , which number five. In other words, there exist multiple parametric configurations (i.e., different θ) that would lead to the same distribution of the observations. So by applying the definition of identifiability, there exist vectors $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$, such that $\theta_1 \neq \theta_2$, but $F_{\mathbf{Z}}(\mathbf{a} : \theta_1) = F_{\mathbf{Z}}(\mathbf{a} : \theta_2)$, for all $\mathbf{a}$. Thus, the model is not identified.

It is possible, however, that certain individual parameters are identifiable. For this model, μ_x is identified since the mean of $\underline{x}$ is equal to the mean of $\underline{X}$. In

order to construct a consistent estimator for the entire vector θ , however, one must specify additional information for the model.

2.2.2 Additional Information Required for Identifiability

Known Parameters

Identifying information could come in the form of known parameters, such as known measurement error variance, σ_{uu} , or reliability coefficient, κ_{xx} , for the measured value of x_i .

If the reliability ratio, $\kappa_{xx} = \sigma_{XX}^{-1}\sigma_{xx}$, defined in Section 2.2.1, is assumed to be known, then the unbiased estimator for the regression coefficient β_1 , in the structural model, is given by

$$\hat{\beta}_1 = \kappa_{xx}^{-1} \hat{\beta}_{1,OLS},$$

where $\hat{\beta}_{1,OLS}$ is the least squares coefficient defined by Equation 2.10. Here, $\hat{\beta}_1$ is referred to as the regression coefficient corrected for attenuation.

Now let us assume σ_{uu} is known. Under the given model, $\mathbf{Z}_i = (Y_i, X_i)$ has a bivariate normal distribution and therefore the sample mean $\bar{\mathbf{Z}} = (\bar{Y}, \bar{X})$ and the sample covariances (m_{YY}, m_{XY}, m_{XX}) form a set of sufficient statistics for parameter estimation. The sample covariances are computed in their usual way, e.g., $m_{XY} = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$. Fuller [24, p. 14] refers to m_{YY} , m_{XY} , and m_{XX} as the maximum likelihood estimators adjusted for degrees of freedom. This is due to the fact that when there are no parametric restrictions on the covariance matrix of $\mathbf{Z}_i$, $n^{-1}(n-1)m_{ZZ}$ is the maximum likelihood estimator of the covariance of $\mathbf{Z}_i$. When the parameter vector is identified, the maximum likelihood estimator will be a function of the sufficient statistics. Recall that the population moments of (Y_i, X_i) under the model defined by 2.6 and 2.7, satisfy

$$\begin{aligned} (\sigma_{YY}, \sigma_{XY}, \sigma_{XX}) &= (\beta_1^2 \sigma_{xx} + \sigma_{ee}, \beta_1 \sigma_{xx}, \sigma_{xx} + \sigma_{uu}) \\ (\mu_Y, \mu_X) &= (\beta_0 + \beta_1 \mu_x, \mu_x). \end{aligned} \tag{2.11}$$

By substituting the sample estimators of the unknown population moments into the left hand side of Equation 2.11, one creates a system of equations that can be solved to obtain the parameter estimates. Doing so results in the following estimators:

$$\begin{aligned}\hat{\beta}_1 &= (m_{XX} - \sigma_{uu})^{-1}m_{XY}, \\ (\hat{\sigma}_{xx}, \hat{\sigma}_{ee}) &= (m_{XX} - \sigma_{uu}, S_{YY} - \hat{\beta}_1 m_{XY}), \\ (\hat{\mu}_x, \hat{\beta}_0) &= (\bar{X}, \bar{Y} - \hat{\beta}_1 \bar{X}),\end{aligned}$$

where $\hat{\sigma}_{xx}$ and $\hat{\sigma}_{ee}$ can be negative with positive probability. Knowing σ_{uu} allows us to construct a one-to-one mapping of the minimal sufficient statistic to the vector $(\hat{\mu}_x, \hat{\sigma}_{xx}, \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_{ee})$. In order for these estimators to be proper estimators, they must lie in the parameter space. So $\hat{\sigma}_{xx}$ and $\hat{\sigma}_{ee}$ must both be nonnegative. These two estimators will both be positive as long as $m_{XX} - \sigma_{uu} > 0$ and $m_{YY} - \hat{\beta}_1 m_{XY} > 0$, or equivalently, $m_{YY}(m_{XX} - \sigma_{uu}) - m_{XY}^2 > 0$.

Two other methods for making use of this type of additional information are described next. First, a method known as *simulation-extrapolation*, or SIMEX, was first introduced by Cook and Stefanski [19] and is also discussed in detail in Carroll et al. [13]. This method is employed when there is some additional information available and is useful in the general ME model [13]. The basic idea behind SIMEX is that in a simulation step, additional independent measurement error is simulated and added to the original measured values in order to create additional data sets with successively more variable values of the independent variable. In the extrapolation step, the trend of the bias induced by the measurement error versus the variance of the additional measurement error is determined and then the trend is extrapolated back to the case of no measurement error to provide a nearly unbiased parameter estimate.

The SIMEX procedure is easiest to understand under the simple linear ME model defined by Equations 2.6 - 2.9. We still assume σ_{uu} is known. First additional, independent measurement error with variance $\lambda_m \sigma_{uu}$, where $0 = \lambda_1 < \lambda_2 < \dots < \lambda_M$, is generated and added to the original data. Cook and Stefanski [19, p. 1317] recommend letting lambda range from zero to two and in most of their examples they use a rather coarse grid, namely, $\lambda \in \{0, 0.5, 1, 1.5, 2\}$. So the total measurement error in the m^{th} data set is $(1 + \lambda_m)\sigma_{uu}$ and the least squares estimator of the slope parameter in the m^{th} data set, $\hat{\beta}_{1,m}$, would consistently estimate $\beta_1 \left(\frac{\sigma_{xx}}{\sigma_{xx} + (1 + \lambda_m)\sigma_{uu}} \right)$. Finally, the problem is thought of as a nonlinear regression problem with $\hat{\beta}_{1,m}$ as the dependent variable and λ_m as the independent variable, having a mean function of the form $G(\lambda) = \left(\frac{\beta_1 \sigma_{xx}}{\sigma_{xx} + (1 + \lambda)\sigma_{uu}} \right)$, $\lambda \geq 0$, and extrapolation back to $\lambda = -1$, the “no-measurement error” case, yields the parameter estimate. See Figure 2-1 for a generic SIMEX plot.

One major drawback to the SIMEX method of bias correction, as noted by Cook and Stefanski [19, p. 1324], is that “in applications to real data, it is generally the case that all assumptions are violated to some extent, and real data are seldom as amenable to analysis as simulated data.” So although their simulation results are encouraging, i.e., from a method which is useful and easily applicable to reduce measurement error bias, impressions taken from these simulations may be consequently optimistic, but one should be wary of problems that arise in applications to real data.

A second general approach to incorporating additional identifying information into a measurement error analysis was first introduced by Prentice [44] for the proportional hazards model and as a general approach by Carroll and Stefanski [15]. It is known as the *regression calibration* method. This method is discussed in detail by Carroll et al. [13]. Regression calibration, like SIMEX, cannot be implemented without some additional information. In fact, these should be viewed

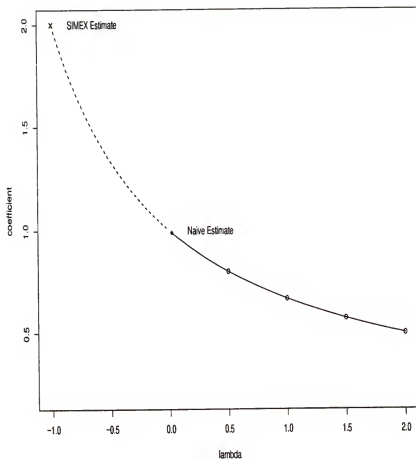


Figure 2-1: SIMEX Plot

A generic SIMEX Plot showing the effect of measurement error of size $(1+\lambda)\sigma_{uu}$ on parameter estimates. Note that the SIMEX Estimate occurs at $\lambda = -1$ and the Naive Estimate occurs at $\lambda = 0$.

as methods of incorporating the additional information that allows parameter identification.

The basic idea of regression calibration is to replace x_i by an estimate of $E(x_i | Z_i, X_i)$, where Z_i are covariates measured without error, and then perform a standard analysis. Note that there are no Z_i covariates in the simple linear ME model, but we include them in this discussion because most existing literature which specifically implements regression calibration includes such error free covariates. One draw back to this method is that estimating $E(x_i | Z_i, X_i)$ often requires information that is specific to the current problem only, and there is no general methodology to estimate this quantity. In the simple linear ME model with known κ_{xx} or σ_{uu} , the regression calibration method produces the usual correction for attenuation. This is easily seen with the present simple linear ME model under the current normality assumptions since it is known that $E(x_i | X_i) = \mu_x + \sigma_{XX}^{-1}\sigma_{xX}(X_i - \mu_x)$. Note that $\bar{X}$ is an unbiased estimate of μ_x and $\sigma_{XX}^{-1}\sigma_{xX} = \sigma_{xx}/(\sigma_{xx} + \sigma_{uu}) = \kappa_{xx}$ is known or can be estimated when κ_{xx} or σ_{uu} is known or estimated. So regression calibration, under the current model, amounts to needing to estimate or have knowledge about κ_{xx} . The resulting estimator using the regression calibration technique under the model defined by 2.6 and 2.7 and with assumptions defined in 2.9 with σ_{uu} known, for example, is $\hat{\beta}_1 = (m_{XX} - \sigma_{uu})^{-1}m_{XY} = \hat{\kappa}_{xx}^{-1}\hat{\beta}_{1,OLS}$. When κ_{xx} is known the regression calibration estimator is $\hat{\beta}_1 = \kappa_{xx}^{-1}\hat{\beta}_{1,OLS}$.

Repeated Measures of True Values

Independent, repeated measures of the truth allows for an independent estimate of σ_{uu} in the simple linear ME model. Suppose there are two replicate observations, then we would have $X_{i1} = x_i + u_{i1}$ and $X_{i2} = x_i + u_{i2}$. So $(X_{i1} - X_{i2}) = (u_{i1} - u_{i2})$ and $S_{X_1-X_2}^2 = \widehat{Var}(u_{i1} - u_{i2}) = 2\hat{\sigma}_{uu}$. Therefore $\hat{\sigma}_{uu} = S_{X_1-X_2}^2/2$. Then a consistent estimate of κ_{xx} , the reliability ratio, is

$\hat{\kappa}_{xx} = (m_{XX} - \hat{\sigma}_{uu})/m_{XX}$, since $\widehat{Var}(X) = \widehat{Var}(x) + \widehat{Var}(u)$ under assumption 2.9, and the resulting estimator, derived from the method of moments, is $\hat{\beta}_{1,OLS}/\hat{\kappa}_{xx}$, the bias adjusted version of $\hat{\beta}_{1,OLS}$. As in the case with known measurement error, this estimator is also the regression calibration estimator. Usually in cases with repeated observations, the replicate means may be thought of as better measure of x than a single observation [13, p. 13]. Therefore these replicate means are used in place of the single observation throughout the analysis when repeated measures are available, i.e., rewrite the model as $Y_i = \beta_0 + \beta_1 x_i + e_i$ and $\bar{X}_i = x_i + \bar{u}_i$.

With an independent estimate of σ_{uu} , Carroll et al. [13] state that either SIMEX or the regression calibration technique can be used to develop estimators for β_1 . The regression calibration estimator is given there as $\hat{\beta}_{1,OLS} m_{XX} / (m_{XX} - \hat{\sigma}_{uu}) = \hat{\kappa}_{xx}^{-1} \hat{\beta}_{1,OLS}$.

Instrumental Variables

Assume, again, that the model defined by 2.6 and 2.7, the simple linear ME model, holds. So, we have

$$\begin{bmatrix} x_i \\ e_i \\ u_i \end{bmatrix} \sim^{iid} N \left(\begin{bmatrix} \mu_x \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & \sigma_{xe} & \sigma_{xu} \\ \sigma_{xe} & \sigma_{ee} & \sigma_{eu} \\ \sigma_{xu} & \sigma_{eu} & \sigma_{uu} \end{bmatrix} \right). \quad (2.12)$$

In this context, the definition of an instrumental variable (IV), given by Carroll et al. [13, p. 107], is a variable that is correlated with the true variable, x_i , but is uncorrelated with the measurement error, u_i , and the model error, e_i . One possible choice of an instrumental variable is a second measure of x_i from a conditionally independent (given x_i) method of measurement (see, for example, [17]), or similarly from a second response variable that is correlated to x_i but independent both of u_i and e_i . Greenland [28] provides a nice introduction and overview of IV's for a non-statistical audience by discussing their role in

epidemiological studies. We shall denote the instrumental variable as W_i . Fuller [24, p. 51] provides a formal definition. According to his definition, under the model defined by 2.6 and 2.7, W_i is an instrumental variable if the following two conditions are met:

$$\begin{aligned} (i) \quad & E \left\{ n^{-1} \sum_{i=1}^n (W_i - \bar{W})(e_i, u_i) \right\} = (0, 0) \\ (ii) \quad & E \left\{ n^{-1} \sum_{i=1}^n (W_i - \bar{W})x_i \right\} \neq 0 \end{aligned} \quad (2.13)$$

where $\bar{W} = n^{-1} \sum_{i=1}^n W_i$. For convenience, let us express the fact that x_i and W_i are related by using a parametric expression. To do this, we shall denote the parameters of the population regression of x_i on W_i by π_2 , for the slope, and π_1 , for the intercept, where

$$\begin{aligned} \pi_2 &= \left[E \left\{ \sum_{i=1}^n (W_i - \bar{W})^2 \right\} \right]^{-1} E \left\{ \sum_{i=1}^n (W_i - \bar{W})(x_i - \bar{x}) \right\}, \\ \pi_1 &= E \{ \bar{x} - \pi_2 \bar{W} \}. \end{aligned}$$

Using these parameters in the regression equation of x_i on W_i , we can write $x_i = \pi_1 + \pi_2 W_i + r_i$, $i = 1, 2, \dots, n$. Here, r_i represents the failure of x_i to be perfectly linearly related to W_i , or simply, the model error in this regression equation. By the least squares regression method used to construct this equation, r_i has zero correlation with W_i , i.e., the “error” term is independent of the “independent” variable. By substituting the fact that $X_i = x_i + u_i$, we have $X_i = \pi_1 + \pi_2 W_i + (r_i + u_i) = \pi_1 + \pi_2 W_i + a_i$. Note that $E \{ \sum_{i=1}^n W_i a_i \} = 0$ by the assumption $E \{ n^{-1} \sum_{i=1}^n (W_i - \bar{W})u_i \} = 0$.

Here $\mu_x = \pi_1 + \pi_2 \mu_W$ and note that we do not assume that σ_{eu} or σ_{xu} is zero. Assuming the IV is normally distributed, along with the assumptions from Equation 2.12, we have that the vector of observed data, (Y_i, X_i, W_i) , is normally distributed with mean $(\mu_Y, \mu_X, \mu_W) = (\beta_0 + \beta_1 \pi_1 + \beta_1 \pi_2 \mu_W, \pi_1 + \pi_2 \mu_W, \mu_W)$ and

variance-covariance matrix

$$\begin{bmatrix} \beta_1^2 \sigma_{xx} + 2\beta_1 \sigma_{xe} + \sigma_{ee} & \beta_1 \sigma_{xx} + \sigma_{xe} + \beta_1 \sigma_{xu} + \sigma_{eu} & \beta_1 \pi_2 \sigma_{WW} \\ \beta_1 \sigma_{xx} + \sigma_{xe} + \beta_1 \sigma_{xu} + \sigma_{eu} & \sigma_{xx} + 2\sigma_{xu} + \sigma_{uu} & \pi_2 \sigma_{WW} \\ \beta_1 \pi_2 \sigma_{WW} & \pi_2 \sigma_{WW} & \sigma_{WW} \end{bmatrix}.$$

Under the model defined by 2.6, 2.7, and 2.12 and in the presence of an instrumental variable, there are 12 independent unknown parameters, namely β_0 , β_1 , μ_W , π_1 , π_2 , σ_{xx} , σ_{xe} , σ_{ee} , σ_{xu} , σ_{eu} , σ_{uu} , and σ_{WW} . This easily can be seen because we have the following assumptions:

$$\begin{bmatrix} x_i \\ e_i \\ u_i \\ W_i \end{bmatrix} \sim^{iid} N \left(\begin{bmatrix} \mu_x \\ 0 \\ 0 \\ \mu_W \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & \sigma_{xe} & \sigma_{xu} & \sigma_{xW} \\ \sigma_{xe} & \sigma_{ee} & \sigma_{eu} & 0 \\ \sigma_{xu} & \sigma_{eu} & \sigma_{uu} & 0 \\ \sigma_{xW} & 0 & 0 & \sigma_{WW} \end{bmatrix} \right).$$

Note, again, that $\mu_x = \pi_1 + \pi_2 \mu_W$ and we assume $\sigma_{xW} = \pi_2 \sigma_{WW} \neq 0$. There are 9 sample statistics that make up the set of minimal sufficient statistics for a sample of n observations. They are from the sample mean vector and sample variance-covariance matrix of (Y_i, X_i, W_i) . We know, therefore, that we cannot hope to estimate all 12 population parameters. Nevertheless, we can develop an estimator for β_1 by noting that the ratio of covariances $\sigma_{xW}^{-1} \sigma_{YW} = (\pi_2 \sigma_{WW})^{-1} \beta_1 \pi_2 \sigma_{WW} = \beta_1$. So, we can estimate β_0 and β_1 by

$$\hat{\beta}_1 = m_{XW}^{-1} m_{YW}, \quad (2.14)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad (2.15)$$

where $\bar{Y}$ and $\bar{X}$ are the sample means and m_{XW} and m_{YW} are the sample covariances. Under the stated assumptions, $\hat{\beta}_0$ and $\hat{\beta}_1$ are consistent estimators, since the sample moments are consistent estimators of the population moments.

Fuller [24, p. 53] provides a theorem and proof concerning the asymptotics of the estimators given in 2.14 and 2.15. Under the given model assumptions, along with the assumptions that σ_{We} and σ_{Wu} are zero, $\sigma_{xW} \neq 0$, and $E\{(W_i - \mu_W)^2 v_i^2\} = \sigma_{WW}\sigma_{vv}$ and $E\{v_i^2(W_i - \mu_W)\} = 0$, where $v_i = e_i - \beta_1 u_i$, then

$$n^{1/2} \begin{bmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{bmatrix} \rightarrow^L N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{vv} + \mu_x^2 V_{22} & -\mu_x V_{22} \\ -\mu_x V_{22} & V_{22} \end{bmatrix} \right),$$

where $V_{22} = \sigma_{xW}^{-2} \sigma_{WW} \sigma_{vv}$.

Carter and Fuller [18] discuss instrumental variable estimation and its properties in the simple ME model with slightly more restrictive assumptions. They assumed that $(x_i, e_i, u_i, W_i)'$, $i = 1, 2, \dots, n$, are distributed as independent drawings from a multivariate normal distribution with a zero mean vector and a covariance matrix, Σ , given by

$$\Sigma = \begin{bmatrix} \sigma_{xx} & 0 & 0 & \sigma_{xW} \\ 0 & \sigma_{ee} & \sigma_{eu} & 0 \\ 0 & \sigma_{eu} & \sigma_{uu} & 0 \\ \sigma_{xW} & 0 & 0 & \sigma_{WW} \end{bmatrix}.$$

They derived restricted maximum likelihood estimators for the models where the error covariance, σ_{eu} , is known to be zero and where σ_{eu} is unknown. The consistency of the restricted maximum likelihood estimator of β_1 was also shown and its asymptotic distribution derived. Details of their work under these assumptions can be found in an Iowa State University dissertation by Carter [16].

Recall in regression calibration an estimate of $E(x_i | Z_i, X_i)$, where Z_i are covariates measured without error (which do not exist in the present model), may be used to replace x_i and then a standard analysis can be run. In the presence of an unbiased IV W_i , (i.e., where $E(x_i | Z_i, X_i) = E(W_i | Z_i, X_i)$, which is the case, for example, when $x_i = \pi_0 + \pi_1 W_i + r_i$ and $\pi_0 = 0$, $\pi_1 = 1$) Carroll et al.

[13] point out that $\hat{E}(W_i \mid Z_i, X_i)$ is obtained from the regression of W_i on Z_i and X_i and is an unbiased estimate of $E(x_i \mid Z_i, X_i)$. When no such unbiased instrument exists, which is usually the case, and one wishes to estimate $E(x_i \mid X_i)$, recall under normality this expectation is $\mu_x + \sigma_{XX}^{-1}\sigma_{xX}(X_i - \mu_x)$, where μ_x is estimated unbiasedly by $\bar{X}$ and one is left with estimation of $\sigma_{XX}^{-1}\sigma_{xX} = \kappa_{xx}$. When an IV is present, it can be shown that an estimate of κ_{xx} may be obtained by regressing W on X . The resulting estimator is $\hat{\kappa}_{xx} = m_{XY}m_{XW}(m_{YW}m_{XX})^{-1}$ which can be seen from the variance-covariance matrix of the observed data since $\sigma_{XY}\sigma_{XW}(\sigma_{YW}\sigma_{XX})^{-1} = \beta_1\sigma_{xx}\pi_2\sigma_{WW}(\beta_1\pi_2\sigma_{WW}(\sigma_{xx} + \sigma_{uu}))^{-1} = \sigma_{xx}(\sigma_{xx} + \sigma_{uu})^{-1}$, when σ_{xe} , σ_{xu} , and σ_{eu} are zero.

SIMEX estimators rely on knowing or estimating the measurement error variance and using this to generate data sets with successively larger measurement error variances and using these new data sets then to extrapolate back to the no measurement error case. Because the SIMEX method uses measurement error variance as the identifying piece of information, SIMEX estimators could be used with instrumental variables by using the IV to estimate σ_{uu} and then using $\hat{\sigma}_{uu}$ in the SIMEX procedure. This, of course, would be unnecessary except in a study to compare these two methods of estimation

2.3 The Multivariable Linear ME Model Under Normality

Now we consider the linear ME model that contains more than one x variable. Let

$$Y_i = \beta_0 + \underline{x}_i' \underline{\beta}_1 + e_i, \quad (2.16)$$

$$\underline{X}_i = \underline{x}_i + \underline{u}_i, \quad i = 1, 2, \dots, n, \quad (2.17)$$

where $\underline{x}_i$ is a k -dimensional column vector, $\underline{\beta}_1$ is a k -dimensional column vector, and the $(k+1)$ -dimensional vectors $\underline{\varepsilon}_i = (e_i, \underline{u}_i)'$ are independent normal $(0, \Sigma_{\varepsilon\varepsilon})$ random vectors. If we assume $\underline{x}_i$ in Equation 2.17 to be normally distributed and

independent of e_i and $\underline{u}_i$, then we have the following:

$$\begin{bmatrix} \underline{x}_i \\ e_i \\ \underline{u}_i \end{bmatrix} \sim^{iid} N \left(\begin{bmatrix} \underline{\mu}_x \\ 0 \\ \underline{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \underline{0} & \underline{0} \\ \underline{0} & \sigma_{ee} & \Sigma_{eu} \\ \underline{0} & \Sigma'_{eu} & \Sigma_{uu} \end{bmatrix} \right). \quad (2.18)$$

It is worth noting here that Σ_{eu} is a $1 \times k$ row vector.

As in the simple ME model, for different estimation purposes, e_i and $\underline{u}_i$ may be assumed to be independent. The vector of observed data, $(Y_i, \underline{X}'_i)$, has a multivariate normal distribution with mean vector $E\{(Y_i, \underline{X}'_i)\} = (\mu_Y, \underline{\mu}'_X) = (\beta_0 + \underline{\mu}'_x \beta_1, \underline{\mu}'_x)$ and variance-covariance matrix

$$Var\{(Y_i, \underline{X}'_i)\} = \begin{bmatrix} \sigma_{YY} & \Sigma_{XY} \\ \Sigma'_{XY} & \Sigma_{XX} \end{bmatrix} = \begin{bmatrix} \beta'_1 \Sigma_{xx} \beta_1 + \sigma_{ee} & (\Sigma_{xx} \beta_1)' + \Sigma_{eu} \\ \Sigma_{xx} \beta_1 + \Sigma'_{eu} & \Sigma_{xx} + \Sigma_{uu} \end{bmatrix}.$$

2.3.1 Non-Identifiability and Bias

In addition to the model defined in Equations 2.16 and 2.17 with assumptions defined in 2.18, let us assume e_i and $\underline{u}_i$ are independent, for all i . Then we have the vector

$$(\underline{x}'_i, e_i, \underline{u}'_i)' \sim^{iid} N \left[(\underline{\mu}'_x, 0, \underline{0})', \text{ block diag}(\Sigma_{xx}, \sigma_{ee}, \Sigma_{uu}) \right]$$

and the variance-covariance matrix for the observed data becomes

$$Var\{(Y_i, \underline{X}'_i)\} = \begin{bmatrix} \sigma_{YY} & \Sigma_{XY} \\ \Sigma'_{XY} & \Sigma_{XX} \end{bmatrix} = \begin{bmatrix} \underline{\beta}'_1 \Sigma_{xx} \beta_1 + \sigma_{ee} & (\Sigma_{xx} \beta_1)' \\ \Sigma_{xx} \beta_1 & \Sigma_{xx} + \Sigma_{uu} \end{bmatrix}.$$

Naively estimating the $\underline{\beta}_1$ parameter vector using ordinary least squares gives the estimator $\hat{\underline{\beta}}_{1,OLS} = S_{XX}^{-1} S_{XY}$ which is biased when $\Sigma_{uu} \neq 0$ because $E(\hat{\underline{\beta}}_{1,OLS}) = (\Sigma_{xx} + \Sigma_{uu})^{-1} \Sigma_{xx} \beta_1$, when $\underline{X}_i$ and Y_i are jointly normal.

The structural ME model defined by Equations 2.16 and 2.17, with assumptions in 2.18, is not identified without additional information, according to Fuller's

definition of identifiability stated in Section 2.2.1. Having the additional information that Σ_{eu} and Σ_{uu} are known, Fuller [24] states that the model is identified because in that case the number of unknown parameters in the vector $\underline{\theta}$, namely $[\underline{\mu}'_x, \beta_0, \underline{\beta}'_1, \sigma_{ee}, \text{vech } \Sigma_{xx}]$, where $\text{vech } A$ of a symmetric matrix A is the vector formed from the distinct elements of A (i.e., $\text{vech } A = (a_{11}, a_{21}, a_{31}, a_{22}, a_{32}, a_{33})$ if A is a 3×3 matrix) is equal to the number of elements in the vector of statistics $[\bar{Y}, \bar{X}', \text{vech } \widehat{\text{Var}}\{(Y_i, \underline{X}'_i)\}]$, where $\widehat{\text{Var}}\{(Y_i, \underline{X}'_i)\}$ is the MLE of $\text{Var}\{(Y_i, \underline{X}'_i)\}$. Therefore, when Σ_{eu} and Σ_{uu} are unknown, the number of unknown parameters in $\underline{\theta}$ exceeds the number of distinct elements in the vector of statistics $[\bar{Y}, \bar{X}', \text{vech } \widehat{\text{Var}}\{(Y_i, \underline{X}'_i)\}]$, meaning that it is not possible to find a unique relationship between $\underline{\theta}$ and the the distribution of the vector $(Y_i, \underline{X}'_i)$. So applying Fuller's definition gives the result that without additional information, the model is not identified.

2.3.2 Additional Information Required

Known Parameters

We shall now see that when some parameters are assumed known, applying likelihood methods to the structural form of model 2.16 and 2.17, gives estimators that are consistent and asymptotically normal. Let $\underline{x}$ be random in the model defined by 2.16 and 2.17, under assumptions defined in 2.18. Assume that Σ_{uu} and Σ_{eu} are both known. Fuller [24] shows that the number of parameters in the vector $\underline{\theta}$ are equal to the number of elements in the vector of sufficient statistics. Therefore, because the sample mean vector and $(n - 1)/n$ times the sample covariance matrix maximize the likelihood, the functional invariance property of the maximum likelihood method gives the maximum likelihood estimators adjusted

for degrees of freedom as

$$\begin{aligned}
 (\hat{\underline{\mu}}'_x, \hat{\beta}_0) &= (\bar{X}', \bar{Y} - \bar{X}'\hat{\beta}_{\underline{1}}), \\
 \hat{\beta}_{\underline{1}} &= (m_{XX} - \Sigma_{uu})^{-1}(m'_{XY} - \Sigma'_{eu}), \\
 \hat{\sigma}_{ee} &= m_{YY} - 2m_{XY}\hat{\beta}_{\underline{1}} + \hat{\beta}'_{\underline{1}}m_{XX}\hat{\beta}_{\underline{1}} + 2\Sigma_{eu}\hat{\beta}_{\underline{1}} - \hat{\beta}'_{\underline{1}}\Sigma_{uu}\hat{\beta}_{\underline{1}}, \\
 \hat{\Sigma}_{xx} &= m_{XX} - \Sigma_{uu},
 \end{aligned}$$

where m represents the unbiased sample covariance matrix. Again, these estimators could result in negative estimates with positive probability, however, the quantities given are proper estimators provided $\hat{\sigma}_{ee}$ is greater than zero and $\hat{\Sigma}_{xx}$ is positive definite.

Carroll et al. [13, p. 95] show that in multiple linear regression the SIMEX estimator with the error variance known results in the usual method of moments estimator when an extrapolant function of the form $G(\lambda, \Gamma) = \gamma_1 + \frac{\gamma_2}{\gamma_3 + \lambda}$, where $\Gamma = (\gamma_1, \gamma_2, \gamma_3)'$, is used.

Repeated Measures

Usually in practice, the matrices Σ_{uu} and Σ_{ue} are unknown, but they can be estimated when repeated measures are observed. If the measurement error covariance matrices are to be estimated, more details must be given about the model defined in 2.16 and 2.17 since e_i may be composed of two parts. The true y_i should be expressed with error in the equation, i.e., $y_i = \underline{x}'_i\beta + q_i$, where q_i are independent $(0, \sigma_{qq})$ random variables that are independent of $\underline{x}_j$ for all i and j . We observe $(Y_i, \underline{X}'_i) = (y_i, \underline{x}'_i) + (w_i, \underline{u}'_i)$, where $(w_i, \underline{u}'_i) = \underline{a}'_i \sim NI(0', \Sigma_{aa})$, and $\underline{a}'_i$ is independent of $(q_j, \underline{x}'_j)$ for all i and j . Fuller [24, p. 106] notes that generally the variance of q_i is not known, however, it is possible to conduct experiments to estimate the covariance matrix of $\underline{a}'_i = (w_i, \underline{u}'_i)$. He also provides the estimator of the coefficient vector, which is analogous to $\hat{\beta}$ in the known parameter case, as $\tilde{\beta} = (m_{XX} - S_{uu})^{-1}(m'_{XY} - S'_{uw})$ where S_{aa} is an unbiased estimator Σ_{aa} , allowing

one to obtain S_{uu} and S_{uw} . Note that $\tilde{\underline{\beta}}$ now contains terms for the intercept and the slope parameters. Further, Fuller gives a consistent estimator of σ_{qq} as $\tilde{\sigma}_{qq} = s_{vv} - (S_{ww} - 2\tilde{\underline{\beta}}'S_{uw} + \tilde{\underline{\beta}}'S_{uu}\tilde{\underline{\beta}})$, where $s_{vv} = (n - k)^{-1} \sum_{i=1}^n (Y_i - \underline{X}_i'\tilde{\underline{\beta}})^2$, and gives a theorem and proof concerning the limiting distribution of the estimators $\tilde{\underline{\beta}}$ and $\tilde{\sigma}_{qq}$. Letting $\tilde{\underline{\theta}}' = (\tilde{\underline{\beta}}', \tilde{\sigma}_{qq})$, the theorem states that $\tilde{\underline{\theta}}$ is consistent for $\underline{\theta}' = (\underline{\beta}', \sigma_{qq})$, has a normal limiting distribution, and provides a formula for its asymptotic variance.

Fuller [23] identifies five cases on different amounts of information available for estimating the covariance matrix of the measurement error. In all cases, the limiting distribution of the resulting estimators, normalized by $n^{1/2}$, is normal. The estimators he provides are consistent for $\underline{\beta}$ and he gives formulae for their asymptotic covariance matrices.

Instrumental Variables

Assume the model defined by 2.16 and 2.17 with the addition of a q -dimensional vector of instrumental variables, $\underline{W}_i$, where the number of instrumental variables is equal to or exceeds the number of variables measured with error, i.e., $n > q \geq k$. Conditions, given by Fuller [24, p. 149], that n and $\underline{W}_i$ must satisfy are that $\sum_{i=1}^n \underline{W}_i \underline{W}_i'$ is nonsingular with probability one, $E\{\underline{W}_i(e_i, \underline{u}_i')\} = (\underline{0}, 0)$, and the rank of $(\sum_{i=1}^n \underline{W}_i \underline{W}_i')^{-1} \sum_{i=1}^n \underline{W}_i \underline{X}_i'$ is k , the dimension of $\underline{x}_i$, with probability one. This specification is to allow for a zero variance for some of the elements of $\underline{X}_i$. An individual element of $\underline{x}_i$ measured without error must be a linear function of $\underline{W}_i$. Therefore, a variable measured without error may serve as the instrumental variable for itself.

Fuller [24] developed the instrumental variable estimator for the coefficient vector in a method similar to that used for the simple linear ME model. The estimator is given by

$$\hat{\underline{\beta}}_1 = (W'X)^{-1}(W'Y),$$

where W is the $q \times n$ matrix of observations on the q instrumental variables. Fuller gave a theorem and proof stating the estimator is consistent and has a normal limiting distribution. A formula is also given for its asymptotic variance.

As previously mentioned, there is no practical use to perform SIMEX estimation with instrumental variables since SIMEX relies on either knowing or estimating the measurement error variance as the required identifying piece of information. To use SIMEX with instrumental variables, an estimate of Σ_{uu} would be obtained after the estimation $\underline{\beta}_1$. Therefore, it is not logical to use SIMEX with IV's.

2.4 Nonlinear ME Models

In this section we discuss ME models which are nonlinear in either $\underline{x}_i$ or $\underline{\beta}$. Fuller [24] says it is conventional to consider ME models to be nonlinear only when the $\underline{\beta}$'s enter the mean function in a nonlinear manner or when the mean function is nonlinear in the explanatory variables measured with error. Let the model be defined by

$$y_i = h(\underline{x}_i; \underline{\beta}) + q_i, \quad (2.19)$$

$$Y_i = y_i + w_i, \quad (2.20)$$

$$\underline{X}_i = \underline{x}_i + \underline{u}_i, \quad (2.21)$$

where $i = 1, 2, \dots, n$ and $h(\cdot)$ is a real valued continuous, nonlinear function. We may combine Equations 2.19 and 2.20 into one equation giving $Y_i = h(\underline{x}_i; \underline{\beta}) + e_i$, where e_i is the sum of both measurement error, w_i , and random equation error, q_i , if they exist, and $\underline{\varepsilon}'_i = (e_i, \underline{u}'_i)$ are independent random vectors having mean $\underline{0}$ and covariance matrix Σ_{ee} . Often it is assumed that the errors are normally distributed. We will primarily be interested in estimating the unknown vector $\underline{\beta}$. This proves, however, to be more challenging in nonlinear ME models compared to linear ME models since, "in most situations it is not possible to obtain an explicit expression for the maximum likelihood estimator of the parameter vector for the nonlinear

model” [24, p. 230]. Fuller [24, p. 226] provides a formal definition of a nonlinear ME model stating the model defined by 2.19, 2.20, and 2.21 is nonlinear if $h(\underline{x}; \underline{\beta})$ is nonlinear in $\underline{x}$ when $\underline{\beta}$ is fixed or if $h(\underline{x}; \underline{\beta})$ is nonlinear in $\underline{\beta}$ when $\underline{x}$ is fixed. We will once again focus mainly on the structural case where the $\underline{x}_i$ are assumed to be random with mean $E(\underline{x}_i) = \underline{\mu}_x$ and variance-covariance matrix Σ_{xx} . If $\underline{x}_i$ are assumed to be independent of the errors, then we have

$$\begin{bmatrix} \underline{x}_i \\ e_i \\ \underline{u}_i \end{bmatrix} \sim F \left(\begin{bmatrix} \underline{\mu}_x \\ 0 \\ \underline{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \underline{0} & \underline{0} \\ \underline{0} & \sigma_{ee} & \Sigma_{eu} \\ \underline{0} & \Sigma'_{eu} & \Sigma_{uu} \end{bmatrix} \right), \quad (2.22)$$

where F might be any distribution assumed for the model. In the next subsection, we discuss the case where F is a multivariate normal distribution.

2.4.1 Normal Theory Models: Non-Identifiability and Bias

This class of ME models are those whose error distributions are independent of their mean function. Griliches and Ringstad [29, p. 370] showed that the bias in the classical least squares estimators is exacerbated when the regression function is nonlinear. Their results are for a specific nonlinear model, but they state that their work “can be viewed, however, as an approximation to the estimation of more general nonlinear models”. They assumed additive measurement error, $X = x + u$, with u and x normally distributed, $E(u) = 0$, $E(xu) = 0$, and parameterized the model such that $\sigma_{XX} = 1$. The authors denoted σ_{uu} by λ . Their model was $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$, nonlinear in the random variable x , with $E(e) = E(xe) = E(x^2 e) = 0$, and all variables were univariate. They showed that the naive ordinary least squares estimate of β_1 , $\hat{\beta}_1$, is biased towards zero by a factor of $(1 - \lambda)$, where in their notation, λ was the fraction of error variance in the total variance in the observed variable, i.e., σ_{uu} . The problem became even more serious for the nonlinear terms, in that the naive ordinary least squares estimate of β_2 , $\hat{\beta}_2$, was biased towards zero by the square of the bias factor of the linear term.

As in the linear ME model, the naive estimator would result from least squares estimation using the observed $\underline{X}$ in place of the latent, unobserved $\underline{x}$.

In the general model, the naive approach is to fit the model $Y_i = h(\underline{X}_i; \underline{\beta}) + e_i$ by classical least squares or by a semiparametric or nonparametric fitting method. Gleser [26] improves upon this naive estimator in a paper from the proceedings of the 1989 AMS-IMS-SIAM Joint Summer Research Conference. He works under the assumptions that $E(e) = E(u) = 0$ and that e , u , and x are mutually statistically independent. In fitting the model, he takes independent and identically distributed (i.i.d.) observations, (Y_i, X_i) , on (Y, X) and therefore (x_i, e_i, u_i) are i.i.d. each with the same distribution as (x, e, u) . In his words, the naive approach “in general, leads to inconsistent estimators with a high degree of asymptotic bias.” [26, p. 99] He points out that consistent estimators have been obtained in special cases, but the general problem of finding consistent and efficient estimators is still unsolved.

2.4.2 Normal Theory Models: Additional Information Required

Known Parameters

Gleser [26] discusses solutions to the problem of parameter estimation where identifying parameters are assumed to be known, a case not realistic in applications, and also the case where they are estimated from the data. His solution is to replace $\underline{x}$ with the best linear predictor, $\hat{\underline{x}}$, of $\underline{x}$ given $\underline{X}$. This amounts to the regression calibration method. He points out that when $\underline{\mu}_x$, Σ_{xx} , and Σ_{uu} are known, $\underline{x}$ and $\underline{u}$ are independent, and $\underline{X}$ is normally distributed

$$E(\underline{x} \mid \underline{X}) = \underline{X}\Lambda + \underline{\mu}_x(I_k - \Lambda), \quad (2.23)$$

where $\Lambda = (\Sigma_{xx} + \Sigma_{uu})^{-1}\Sigma_{xx}$ is the reliability matrix, I is the $k \times k$ identity matrix, and k is the dimension of $\underline{X}$. Of course, in practice, some of these quantities may not be known. The equality in 2.23 does not hold when $\underline{X}$ is not normally

distributed, but the right hand side of Equation 2.23 is still the best linear (in $\underline{X}$) predictor of $\underline{x}$ under mean square error [26]. So when these parameters are known, Gleser suggests fitting the model $Y = h(\underline{X}\Lambda + \underline{\mu}_x(I_k - \Lambda); \underline{\beta}) + e = h(E(\underline{x} | \underline{X}); \underline{\beta}) + e$.

An iterative estimation procedure for coefficients of nonlinear “functional” relations was proposed by Wolter and Fuller [65] and assumes a known covariance matrix of $(e_i, \underline{u}_i')$, i.e., Σ_{eu} , or minimally its order is known. Their estimator is a modification of the maximum likelihood estimator for the nonlinear model with normal measurement error. Assuming $(e_i, \underline{u}_i')$ are independent normal $(0, \Sigma_{eu})$ random variables, the maximum likelihood estimators are those values of $\underline{x}$ and $\underline{\beta}$ that minimize the sum of squares

$$\sum_{i=1}^n q(\underline{\beta}, \underline{x}_i; Y_i, \underline{X}_i) = \sum_{i=1}^n \{Y_i - h(\underline{x}_i; \underline{\beta}), \underline{X}_i - \underline{x}_i\}' \Sigma_{eu}^{-1} \{Y_i - h(\underline{x}_i; \underline{\beta}), \underline{X}_i - \underline{x}_i\}.$$

Under their nonlinear model it is not possible to derive an explicit expression for the MLE of $\underline{\beta}$ but Wolter and Fuller do develop two estimators under slightly different assumptions of the order of Σ_{eu} and provide theorems for the limiting distributions of these iteratively defined estimators.

Repeated Measures

In Gleser’s solution mentioned above, when $\underline{\mu}_x$ and/or Λ are unknown, consistent estimators $\hat{\underline{\mu}}$ and $\hat{\Lambda}$ can be used as substitutes. Such consistent estimators can be obtained from the data, $\underline{X}_1, \dots, \underline{X}_n$, and information from prior calibration studies on $(\underline{x}, \underline{X})$ [26]. Using repeated measures, an estimate of Σ_{uu} can be obtained as previously mentioned in the sections on repeated measures under the linear ME models, and $\hat{\Sigma}_{uu}$ can then be used to estimate Λ by $\hat{\Lambda} = S_{XX}^{-1}(S_{XX} - \hat{\Sigma}_{uu})$. In such cases, the model $Y = h(\hat{\underline{x}}; \underline{\beta}) + e$, where $\hat{\underline{x}} = \underline{X}\hat{\Lambda} + \hat{\underline{\mu}}_x(I_k - \hat{\Lambda})$, should be fitted to estimate $\underline{\beta}$. Gleser [26] also points out that a better substitution for $h(\underline{x}; \underline{\beta})$ may be obtained from a Taylor series expansion of $h(\underline{x}; \underline{\beta})$ about $\hat{\underline{x}}$. He proved consistency and asymptotic distributional properties for his estimator.

Instrumental Variables

In addition to the model defined by 2.19, 2.20, and 2.21, assume there are observations available on an instrumental variable, $\underline{W}_i$. In nonlinear models, IV's must still satisfy the same requirements as those stated in Subsection 2.2.2 for the simple linear ME model. That is, IV's must be correlated with $\underline{x}_i$, independent of the measurement error in $\underline{x}_i$, i.e., independent of $\underline{u}_i$, and must also be surrogate, i.e., $\underline{W}_i$ must be independent of Y_i given $\underline{x}_i$ (which is equivalent to saying independent of model error, q_i).

Amemiya [5] gives the IV estimator and its asymptotic properties for the nonlinear “functional” errors in variables model that is nonlinear in $\underline{x}_i$. His model is $y_i^0 = h(\underline{x}_i^0; \underline{\beta}^0)$ with no model error, q_i , and he assumes the measurement errors, $\underline{\varepsilon}_i' = (e_i, \underline{u}_i')$, to be independently distributed with zero mean and unknown covariance matrix $\Sigma_{\varepsilon\varepsilon}$. Because he is working with the functional model, the $\underline{x}_i^0$ are assumed to be fixed constants and therefore have no distributional assumptions. Amemiya defines the instrumental variable estimator $\hat{\underline{\beta}}_{IV}$ of $\underline{\beta}^0$ as the value of $\underline{\beta}$ in the parameter space that minimizes

$$Q(\underline{\beta}) = \left\{ n^{-1} \sum_{i=1}^n [Y_i - h(\underline{X}_i; \underline{\beta})] \underline{W}_i \right\} \times \left\{ n^{-1} \sum_{i=1}^n \underline{W}_i' [Y_i - h(\underline{X}_i; \underline{\beta})] \right\},$$

which is the squared norm of the sample covariance between $\underline{W}_i$ and $Y_i - h(\underline{X}_i; \underline{\beta})$, where the observed $\underline{X}_i$ replaced the unobservable $\underline{x}_i^0$. The motivation behind using $Q(\underline{\beta})$ is that if $\underline{W}_i$ is uncorrelated with measurement errors $\underline{\varepsilon}_i$, then $Y_i - h(\underline{x}_i^0; \underline{\beta})$ is uncorrelated with $\underline{W}_i$ when $\underline{\beta} = \underline{\beta}^0$. Therefore to obtain an estimator of $\underline{\beta}^0$ minimize $Q(\underline{\beta})$. It is noted that this estimator is equivalent to the nonlinear two-stage least squares estimator. Theorems and proofs are provided for the asymptotics of his estimator. It is stated that $plim_{n \rightarrow \infty} \hat{\underline{\beta}}_{IV} = \underline{\beta}^0$ and that $n^{1/2} (\hat{\underline{\beta}}_{IV} - \underline{\beta}^0) \rightarrow^L N(\underline{0}, \underline{V})$, where a formula for $\underline{V}$, the covariance matrix of the limiting distribution of the estimator, is given.

Not much research has been done on structural nonlinear ME models. Buzas [9], however, does assume a structural model in using a regression calibration approach with IV's to construct unbiased score equations in nonlinear measurement error models. His method, however, requires the use of an unbiased estimate of the true, unknown variable.

SIMEX estimation uses knowledge about the measurement error variance as it's identifying piece of information, and therefore, this type of bias adjusting estimation is not used with instrumental variables in nonlinear ME models, except perhaps for comparison between methods.

2.4.3 Exponential Family Models: Non-Identifiability and Bias

In this class of ME models are those whose response distributions can be a function of the mean. A generalized linear model, or GLM, is one type of model that shall be included in this section. See McCullagh and Nelder [40] for further details of the GLM. The observed, dependent variable, Y_i , is linked to the explanatory variables, $\underline{x}_i$, as follows:

$$E(Y_i | \underline{x}_i) = \mu_i,$$

where μ_i is called the systematic component of Y_i . The dependence of μ_i on $\underline{x}_i$ is assumed to be determined by

$$g(\mu_i) = \eta_i,$$

for a monotone specification of $g(\cdot)$, the link function, where $\eta_i = \underline{x}_i' \beta$. The distribution of Y_i given $\underline{x}_i$ may be in the general exponential family, in which case

$$f_{Y_i|x}(y_i | \underline{x}_i) = \exp[\{y_i \theta_i - b(\theta_i)\}/a_i(\phi) + c(y_i, \phi)],$$

where θ is the canonical parameter, $E(Y_i | \underline{x}_i) = b'(\theta_i) = \mu_i$, $Var(Y_i | \underline{x}_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)V(\mu_i)$, where $V(\cdot)$ is the variance function depending on μ_i only,

and ϕ , also denoted as σ^2 , is called the scale or dispersion parameter. Commonly, $a_i(\phi) = \phi/w_i$, where w_i are known, weights.

Stefanski [55] discusses the bias in estimating $\underline{\beta}$ using an M-estimator. An M-estimator provides a consistent solution when solving an unbiased estimating equation in the absence of measurement error. M-estimators are often used to estimate parameters in GLM's. One lets $\psi_i(\cdot)$, $i = 1, 2, \dots, n$, be functions of Y_i , $\underline{x}_i$, and $\underline{\beta}$, which has values in a space that is of the same dimension as that of $\underline{\beta}$. An *M-estimator* is a solution to an *estimating equation* (which is *unbiased* if it has expected value 0), which has the form

$$0 = n^{-1} \sum_{i=1}^n \psi_i(Y_i, \underline{x}_i, \underline{\beta}).$$

In practice, the $\psi_i(\cdot)$'s are not chosen, but rather these functions are defined through the choice of an estimator, e.g., maximum likelihood or least squares. (See, for example, [13, 57] for more on M-estimators.) Stefanski [55] defines an M-estimator, $\tilde{\underline{\beta}}$, for an unknown $\underline{\beta}_0$ where

$$\sum_{i=1}^n \psi_i(\underline{z}_i, \tilde{\underline{\beta}}) = 0. \quad (2.24)$$

Here $\underline{z}_i = (y_i, \underline{x}'_i)$, $i = 1, 2, \dots, n$, are independent random vectors such that $E\{\psi_i(\underline{z}_i, \underline{\beta}_0)\} = 0$, where $\psi(\cdot, \cdot)$ and $\underline{\beta}_0$ are of the same dimension. In Stefanski's notation, the observable data consist of the pairs $(\underline{y}_i, \underline{X}_i)$, where $\underline{X}_i = \underline{x}_i + \sigma \underline{v}_i$, $E(\underline{v}_i | y_i, \underline{x}_i) = 0$, $E(\underline{v}_i \underline{v}'_i | y_i, \underline{x}_i) = \Omega(y_i, \underline{x}_i)$, and σ is a scalar. In order for this error structure to result in the usual additive error structure previously discussed, i.e., $\underline{X}_i = \underline{x}_i + \underline{u}_i$, let $\underline{v}_i = \underline{u}_i/\sigma$. This error structure allows for dependence between the measurement error and the vector $(y, \underline{x}')$ and can accommodate situations when the measurement error may be multiplicative or heteroscedastic. A naive estimator is obtained by using $\underline{Z}_i = (y_i, \underline{X}'_i)$ instead of $\underline{z}_i$ in Equation 2.24. According to Stefanski [55], this estimator, denoted by $\hat{\underline{\beta}}$, converges not to

$\underline{\beta}_0$ but to $\underline{\beta}(\sigma)$ which satisfies $\lim n^{-1} \sum_{i=1}^n E\{\psi_i(y_i, \underline{X}_i, \underline{\beta}(\sigma))\} = 0$. (Note that $\underline{\beta}_0 = \underline{\beta}(\sigma = 0)$.) He notes that generally the value that $\hat{\underline{\beta}}$ converges to, i.e., $\underline{\beta}(\sigma)$, will not be equal to the true parameter value, $\underline{\beta}_0$, and thus $\hat{\underline{\beta}}$, in most cases, is asymptotically biased. The bias can be assessed by the relation between $\underline{\beta}(\sigma)$ and $\underline{\beta}_0$ which Stefanski [55, p. 585] derives from a Maclaurin series expansion. The resulting relation, for σ near zero, is

$$\underline{\beta}_0 = \underline{\beta}(\sigma) + \frac{1}{2}\sigma^2 A^{-1} E\{\psi_{xx}(y, \underline{x}, \underline{\beta}_0)\Omega(y, \underline{x})\} + o(\sigma^2), \quad (2.25)$$

where $A = \lim n^{-1} \sum_{i=1}^n E\left\{\left[\frac{\partial \psi(\underline{x}, \underline{\beta})}{\partial \underline{\beta}}\right]_{\underline{\beta}=\underline{\beta}_0}\right\}$ and ψ_{xx} refers to the second derivative of the second term in $\psi(\cdot, \cdot, \cdot)$ with respect to σ . Unfortunately, Equation 2.25 is a function of the unknown $\underline{\beta}_0$ and the true x 's. Stefanski provides an estimator of this relation which is his bias adjusting estimator. It further adjusts the bias at each step, although he does not pursue higher order corrections. His method requires knowledge or estimability of the measurement error variance, as well as σ in his definition of error structure, and shall be discussed in future sections.

2.4.4 Exponential Family Models: Addition Information Required

Known Parameters

Armstrong [6] provides a method for estimating $\underline{\beta}$ when the measurement error variance is known or estimated, but requires an estimate of $E(\underline{x} \mid \underline{X})$. He applies the Generalized Linear Interactive Modeling, or GLIM, software package, used to perform iteratively reweighted least squares, to transformed data with adjusted model specifications. He provides examples where the distribution of measurement errors, namely, the measurement error variances, are assumed known and states that the same method can be used when the measurement error variance is estimated from replicated measurements. We introduce the procedure here and mention it again when the measurement error variance is not assumed to be known, but rather is estimated from repeated observations. First, the distributional form

of Y given $\underline{x}$ in the original GLM is specified for Y given $\underline{X}$. Next a transformation $t(\underline{X})$ of $\underline{X}$ is chosen, and an inverse link function g^{*-1} , possibly different from the original inverse link function g^{-1} , is found such that

$$E(Y \mid \underline{X}) = \mu^* = g^{*-1}(t(\underline{X})'\underline{\beta}),$$

at least as an approximation. The new g^{*-1} replaces g^{-1} as the inverse link and $t(\underline{X})$, the transformed data, is used as the explanatory variable. An example he provides is the case of GLM's with the identity link. In this case $E(Y \mid \underline{x}) = \mu = \underline{x}'\underline{\beta}$ and hence $E(Y \mid \underline{X}) = E\{E(Y \mid \underline{x}) \mid \underline{X}\} = E(\underline{x}'\underline{\beta} \mid \underline{X}) = \{E(\underline{x} \mid \underline{X})'\}\underline{\beta}$. So here we see $g^{*-1} = g^{-1} = \text{identity}$ and $t(\underline{X}) = E(\underline{x} \mid \underline{X})$, the posterior mean of $\underline{x}$ given $\underline{X}$, which is the piece used to replace the unobservable $\underline{x}$ in the regression calibration method. Finally, apply iteratively reweighted least squares where the iterative reweighting is redefined so as to match the correct variance, $\text{Var}(Y \mid \underline{X})$. Armstrong suggests iteratively specifying a new prior weight w^* as

$$w^* = \frac{\phi\tau^2(\mu^*)}{(\phi/w)E(\tau^2(\mu) \mid \underline{X}) + \text{Var}(\mu \mid \underline{X})},$$

where μ , w , ϕ , and $\tau^2(\mu_i) = b''(\theta_i)$ are the mean, prior weight, scale parameter, and variance function defined for the original GLM for Y given $\underline{x}$, respectively. He provides a theorem and proof stating the above procedure produces maximum quasi-likelihood estimates, henceforth MQLE, for $\underline{\beta}$. Armstrong [6, p. 535] also points out that when $f(Y \mid \underline{X})$ is in the exponential family the MQLE's are MLE's. Further, he notes that the MQLE's and their estimated variances are asymptotically unbiased.

As mentioned before, Stefanski [55] provides a bias adjusting estimator when the error variance is known or an estimate of it is available. By the authors admission, needing this type of information is common, but “does restrict the applicability of the results” [55, p. 584]. Having an initial naive estimate, $\hat{\underline{\beta}}$, the

bias adjusting estimate analogous to Equation 2.25 is

$$\hat{\underline{\beta}}_c(\sigma) = \hat{\underline{\beta}} + \frac{1}{2}\sigma^2 \left\{ \sum_{i=1}^n \psi_{\hat{\underline{\beta}}}(y_i, \underline{X}_i, \hat{\underline{\beta}}) \right\}^{-1} \left\{ \sum_{i=1}^n \psi_{XX}(y_i, \underline{X}_i, \hat{\underline{\beta}}) \Omega(y_i, \underline{X}_i) \right\}, \quad (2.26)$$

where σ^2 is a scalar defined in his error structure in Section 2.4.3, $\psi_{\hat{\underline{\beta}}}$ refers to the first derivative of the function $\psi(., ., .)$ with respect to the third argument, and ψ_{XX} refers to the second derivative of the second argument of $\psi(., ., .)$. He provides asymptotic results for the naive estimator as well as his bias adjusting estimator and provides a small Monte Carlo study comparing the two estimators.

A recent publication by Aitkin and Rocci [2] uses the EM algorithm with the assumption of known measurement error variances to estimate the parameters in GLM's with continuous measurement error in the independent variables .

Repeated Measures

When the measurement error variance, Σ_{uu} , is not assumed to be known it can be estimated from repeated measures. We have mentioned in earlier sections on repeated measures ways of estimating Σ_{uu} . Armstrong [6] provides a method for estimating $\underline{\beta}$ from the data that provides MQLE's when the measurement error variance is estimated from replicated observations.

Stefanski [55] provides a bias adjusting estimator when the error variance is known or can be estimated. He performs a Monte Carlo study of the naive estimator with his bias adjusting estimator (given in Equation 2.26) in both cases when σ^2 , a scalar defined in his error variance structure, is known and when it is estimated through replicated measurements. He states there is no significant differences between the two procedures and actually only reports the results from the method when σ^2 is estimated.

For binary regression models, Carroll et al. [14] use replications to estimate Σ_{uu} in constructing structural maximum likelihood estimates.

In a logistic regression example, Carroll et al. [13] use the regression calibration technique with an estimate of the error variance to assess the effect of saturated fat on the risk of developing breast cancer, adjusted for other variables. Saturated fat is measured via a 24 hour recall of food intake and “is measured with considerable error” and bias [13, p. 43]. They apply the regression calibration technique and compare the naive estimator ignoring measurement error with the estimator using the regression calibration method. The regression calibration method did successfully correct for attenuation and this is seen in a comparison of the two parameter estimates. The naive estimate of the saturated fat coefficient was -0.97 and that from the regression calibration technique was -4.67.

Carroll et al. [13] also compare the naive estimator, the regression calibration estimator, and the SIMEX estimator in another logistic regression problem where the error variance was estimated. In this study, the outcome indicates the occurrence of coronary heart disease based on covariates measured without error and systolic blood pressure, the covariate which is measured with error. They provide the results from all three estimation methods as well as compare estimated standard errors of the estimates, also calculated via different methods, e.g., sandwich and information standard errors. They found that the SIMEX and regression calibration estimators adjust for the bias in comparison to the naive estimator.

Instrumental Variables

Carroll et al. [13] use an instrumental variable approach in the GLM setting which is closely related to regression calibration. They propose three IV estimators and, because these estimators rely on numerous regression parameters, they use the notation employed by Stefanski and Buzas [58] to indicate different parameters from regressing one variable on others. This new notation requires underlining variables. Therefore, it should be understood that in the current

section, appropriate variables are still vector valued, but their underline has been removed to allow for this new notation. For example, $\beta_{Y|\underline{1}Zx}$ is the coefficient of 1, i.e., the intercept, in the generalized linear regression of Y on 1, Z , and x , where Z are covariates measured without error; $\beta'_{Y|\underline{1}Zx}$ is the coefficient for Z in the regression of Y on 1, Z , and x . Notice that because of this notation the underbars indicating vectors must be dropped. With this notation, it is easy to represent subsets of coefficient vectors, e.g., $\beta'_{Y|\underline{1}Zx} = (\beta_{Y|\underline{1}Zx}, \beta'_{Y|\underline{1}Zx})$. The form of the GLM used by Carroll et al. [13] is given by:

$$\begin{aligned} E(Y \mid Z, x) &= b'(\beta_{Y|\underline{1}Zx} + Z'\beta_{Y|\underline{1}Zx} + x'\beta_{Y|\underline{1}Zx}), \\ \text{Var}(Y \mid Z, x) &= \phi V(\beta_{Y|\underline{1}Zx} + Z'\beta_{Y|\underline{1}Zx} + x'\beta_{Y|\underline{1}Zx}), \end{aligned}$$

where $V(\cdot)$ is the variance function and ϕ is the dispersion parameter. Writing the model in this form means that Carroll et al. restrict their methods to GLM's with canonical link functions. This proves to be the first restriction to their methods under GLM's. Further, they use composite vectors to aid in the estimation algorithms where $\tilde{x} = (1, Z', x')'$, $\tilde{X} = (1, Z', X')'$, and $\tilde{W} = (1, Z', W')'$. So we may define $\beta_{Y|\underline{\tilde{x}}} = (\beta_{Y|\underline{1}Zx}, \beta'_{Y|\underline{1}Zx}, \beta'_{Y|\underline{1}Zx})'$ and then the basic model may be written

$$\begin{aligned} E(Y \mid \tilde{x}) &= b'(\tilde{x}'\beta_{Y|\underline{\tilde{x}}}), \\ \text{Var}(Y \mid \tilde{x}) &= \phi V(\tilde{x}'\beta_{Y|\underline{\tilde{x}}}). \end{aligned}$$

A key assumption to their methods, which is yet another restriction, is that the regression of x on (Z, W, X) is approximately linear, i.e.,

$$E(x \mid Z, W, X) = \beta_{x|\underline{1}ZW\bar{X}} + Z'\beta_{x|\underline{1}ZW\bar{X}} + W'\beta_{x|\underline{1}ZW\bar{X}} + X'\beta_{x|\underline{1}ZW\bar{X}}. \quad (2.27)$$

Further, they assume that

$$E(x - X \mid Z, x, W) = 0 \quad (2.28)$$

$$E(Y \mid Z, W, X) = E\{E(Y \mid Z, x) \mid Z, W, X\}. \quad (2.29)$$

Note that assumptions 2.27 and 2.28 imply that $E(x \mid Z, W) = E(X \mid Z, W)$ and that $\beta_{X|\widetilde{W}} = \beta_{x|\widetilde{W}}$. First, consider the regression of Y on $\widetilde{W}$. Then we have $E(Y \mid \widetilde{W}) = b'\{E(\tilde{x} \mid \widetilde{W})'\beta_{Y|\tilde{x}}\} = b'(\widetilde{W}\beta_{\tilde{x}|\widetilde{W}}\beta_{Y|\tilde{x}})$. Using the critical assumption that $\beta_{\tilde{x}|\widetilde{W}} = \beta_{x|\widetilde{W}}$, it follows that

$$\beta_{Y|\widetilde{W}} = \beta_{\tilde{x}|\widetilde{W}}\beta_{Y|\tilde{x}}. \quad (2.30)$$

For their first estimator, Carroll et al. [13] suggest starting with the multivariate regression of $\tilde{X}$ on $\widetilde{W}$ to obtain $\hat{\beta}_{\tilde{x}|\widetilde{W}}$. This is used to obtain their estimate of $E(\tilde{x} \mid \widetilde{W})$; the piece required for this form of regression calibration. Then, to obtain an estimator of $\beta_{Y|\tilde{x}}$, use a generalized linear regression of Y on the predicted values of $\tilde{x}$, i.e., $\widetilde{W}\hat{\beta}_{\tilde{x}|\widetilde{W}}$. They denote this estimator as $\hat{\beta}_{Y|\tilde{x}}^{IV1, RC}$; the first instrumental variable estimator using regression calibration.

Their second means of utilizing the regression calibration approximation works directly from the relation in Equation 2.30. For a fixed nonsingular matrix M_1 , they define $\hat{\beta}_{\tilde{x}|\widetilde{W}}^{(M_1)} = (\hat{\beta}'_{\tilde{x}|\widetilde{W}}M_1\hat{\beta}_{\tilde{x}|\widetilde{W}})^{-1}\hat{\beta}'_{\tilde{x}|\widetilde{W}}M_1$. Thus their second estimator is $\hat{\beta}_{Y|\tilde{x}}^{IV1, (M_1)} = \hat{\beta}_{\tilde{x}|\widetilde{W}}^{(M_1)}\hat{\beta}_{Y|\widetilde{W}}$, where $\hat{\beta}_{Y|\widetilde{W}}$ is the estimated regression coefficient from fitting the generalized model to the $(Y, \widetilde{W})$ data. They point out that when W and X are of the same dimension, this second estimator does not depend on M_1 and thus is identical to their first estimator. However, they note that when the number of instruments is greater than the number of variables measured with error, the choice of M_1 matters. Carroll et al. [13, p. 117-119] derive an estimate of M_1 that minimizes the asymptotic variance of $\hat{\beta}_{Y|\tilde{x}}^{IV1, (M_1)}$.

Their final estimator makes use of the fact the X and W are both surrogates. The derivation of this estimator is involved [13, p. 114-115], and will be omitted here. In defining this estimator, let $\dim(Z)$ be the number of components of Z , the covariates measured without error. Further define $\beta_{Y|\widetilde{W}\widetilde{X}} = \beta_{Y|1ZW\widetilde{X}}$ and $\beta_{Y|\widetilde{W}\widetilde{X}} = (0_{1 \times d}, \beta'_{Y|1ZW\widetilde{X}})'$, where $d = 1 + \dim(Z)$. Then, for a given matrix M_2 , their final IV estimator is $\widehat{\beta}_{Y|\widetilde{X}}^{IV2, (M_2)} = \widehat{\beta}_{\widetilde{X}|\widetilde{W}}^{-(M_2)} (\widehat{\beta}_{Y|\widetilde{W}\widetilde{X}} + \widehat{\beta}_{\widetilde{X}|\widetilde{W}} \widehat{\beta}_{Y|\widetilde{W}\widetilde{X}})$. Again, they [13, p. 119-120] derived an estimate of M_2 that minimizes the asymptotic variance of $\widehat{\beta}_{Y|\widetilde{X}}^{IV2, (M_2)}$ for the case $\dim(W) > \dim(X)$.

They compared these three IV estimators along with the naive estimator in a logistic regression example where the outcome is an indicator of coronary heart disease and systolic blood pressure is the covariate measured with error. A second, independent blood pressure measurement taken from a second examination serves as their instrumental variable. In their first application of the example, W and X are of the same dimension so their first two estimators are equivalent. In the second application, W is taken as a two-dimensional variate from two measurements of blood pressure taken at the second examination, and therefore all three estimates are computed. In all cases their estimators aided in correcting for the attenuation when compared to the naive estimate. The authors also provided asymptotic distribution approximations for their estimators.

For further references on the use of instrumental variables in the exponential family, see, for example, Buzas and Stefanski [10] for the use of IV's in generalized linear models, also, under canonical links, Thoresen and Laake [62] for the use of IV's in a factor analysis approach in logistic regression, and Buzas and Stefanski [11] for the use of IV's in a probit ME model.

As mentioned in previous sections on instrumental variable estimation, SIMEX estimators are not used with instrumental variables since this simulation based method of estimation relies on knowledge of the measurement error variance for

estimation. The rare, and not very practical, exception would be to compare an IV estimate to a SIMEX estimate.

In the following chapter we define a general model that can be applied to most ME models. We call the model the Generalized Simple Measurement Error model. In the remainder of the dissertation we develop an estimation method for its parameters and means of making inferences on them. Further, we show how many known ME model problems fit into the Generalized Simple Measurement Error model framework and can be solved using its methods.

CHAPTER 3 GENERALIZED SIMPLE M.E. MODEL WITH INSTRUMENTAL VARIABLES

3.1 The Model

We consider a general IV solution to nonlinear measurement error problems assuming no error free covariates and at least one instrumental variable. We consider the case of p manifest, or observable, random variables that are related to a single x , the unobservable truth. To allow full generality of our model we must allow for both categorical and non-categorical manifest variables (outcome variable, observed value of x , and IV's). We start by looking at the outcome manifest variable(s). Although the model will include a total of p manifest variables, we start by discussing the outcome(s) of interest, which will be denoted with subscript r^* . So without loss of generality, consider the outcome variable(s), $\underline{Y}_{r^*}$, to be any one of the p manifest variables, which are denoted $\underline{Y}_r$, $r = 1, 2, \dots, p$. We let $\underline{Y}_{r^*}$ given x be a p_{r^*} -dimensional column vector representing the primary outcome variables of interest (previously denoted by a univariate Y in Chapter 2). Assume the distribution of $\underline{Y}_{r^*}$ is known and for the i th individual, $i = 1, 2, \dots, n$, we observe $\underline{Y}_{ir^*}$. Let $\underline{\theta}_{r^*} = (\theta_{r^*1}, \theta_{r^*2}, \dots, \theta_{r^*k_{r^*}})' \in \Theta_{r^*}$ be a $k_{r^*} \times 1$ vector of parameters defining the conditional distribution of $\underline{Y}_{r^*}$. We shall suppress the notation for individuals, i.e., i , for the remainder of this dissertation, except in instances where it may be required.

Note, for example, that $\underline{\theta}_{r^*}$ might be the canonical parameter if $\underline{Y}_{r^*}$ is a scalar random variables with a distribution from the single parameter exponential family, or it might represent a vector of conditional distribution parameters in the case of, say, a multivariate regression for $\underline{Y}_{r^*}$, in which case $\underline{Y}_{r^*} = (Y_{r^*1}, Y_{r^*2}, \dots, Y_{r^*p_{r^*}})'$,

where these variables are conditionally dependent. Further, $\underline{\theta}_{r^*}$ could be a vector of multinomial cell probabilities in the case of a categorical response.

For the outcome vector, $\underline{Y}_{r^*}$, assume there exists a $k_{r^*} \times 1$ vector of parameters of interest that we are interested in modeling, which is denoted by $\underline{\gamma}_{r^*} \in \Gamma_{r^*}$. To do so, we specify the models

$$\begin{aligned}\gamma_{r^*1} &= h_{r^*1}(x; \underline{\beta}_{r^*1}) \\ \gamma_{r^*2} &= h_{r^*2}(x; \underline{\beta}_{r^*2}) \\ &\vdots \\ \gamma_{r^*k_{r^*}} &= h_{r^*k_{r^*}}(x; \underline{\beta}_{r^*k_{r^*}}),\end{aligned}\tag{3.1}$$

or, in vector notation $\underline{\gamma}_{r^*} = \underline{h}_{r^*}(x; \underline{\beta}_{r^*})$, where $\underline{\beta}_{r^*}$ is the $l_{r^*} \times 1$ vector denoting the concatenation of the distinct elements of $\underline{\beta}_{r^*1}, \underline{\beta}_{r^*2}, \dots, \underline{\beta}_{r^*k_{r^*}}$. This is the vector of *model parameters*, i.e., those which we intend to estimate, relating $\underline{Y}_{r^*}$ to x .

Assume that the elements of

$$\underline{h}_{r^*}(x; \underline{\beta}_{r^*}) = \left(h_{r^*1}(x; \underline{\beta}_{r^*1}), h_{r^*2}(x; \underline{\beta}_{r^*2}), \dots, h_{r^*k_{r^*}}(x; \underline{\beta}_{r^*k_{r^*}}) \right)'$$

are differentiable with respect to the elements of $\underline{\beta}_{r^*}$.

Further assume that, in terms of the original k_r parameters of the conditional distribution of $\underline{Y}_{r^*}$

$$\begin{aligned}\gamma_{r^*1} &= T_{r^*1}(\underline{\theta}_{r^*}) \\ \gamma_{r^*2} &= T_{r^*2}(\underline{\theta}_{r^*}) \\ &\vdots \\ \gamma_{r^*k_{r^*}} &= T_{r^*k_{r^*}}(\underline{\theta}_{r^*}),\end{aligned}\tag{3.2}$$

and that the vector of functions $(T_{r^*1}(\underline{\theta}_{r^*}), T_{r^*2}(\underline{\theta}_{r^*}), \dots, T_{r^*k_{r^*}}(\underline{\theta}_{r^*}))'$ define a one-to-one mapping of Θ_{r^*} onto Γ_{r^*} . We can express Equation 3.2 in vector notation

as $\underline{\gamma}_{r^*} = \underline{T}_{r^*}(\underline{\theta}_{r^*})$. Since these functions together define a one-to-one and onto mapping, there exists an inverse mapping defined by a vector of functions and denoted $\underline{T}_{r^*}^{-1}(\underline{\gamma}_{r^*}) = (T_{r^*1}^{-1}(\gamma_{r^*1}), T_{r^*2}^{-1}(\gamma_{r^*2}), \dots, T_{r^*k_{r^*}}^{-1}(\gamma_{r^*k_{r^*}}))'$. Given the model assumptions in Equation 3.1 we may write $\underline{\theta}_{r^*}$ as

$$\begin{aligned} \theta_{r^*1} &= T_{r^*1}^{-1}(h_{r^*}(x; \underline{\beta}_{r^*})) \\ \theta_{r^*2} &= T_{r^*2}^{-1}(h_{r^*}(x; \underline{\beta}_{r^*})) \\ &\vdots \\ \theta_{r^*k_{r^*}} &= T_{r^*k_{r^*}}^{-1}(h_{r^*}(x; \underline{\beta}_{r^*})), \end{aligned} \quad (3.3)$$

Note that the $T_{r^*j}^{-1}(\cdot)$ functions, $j = 1, 2, \dots, k_{r^*}$, are not inverse functions, but together define the inverse mapping from Γ_{r^*} onto Θ_{r^*} . The elements of the vector $\underline{T}_{r^*}^{-1}(h_{r^*}(x; \underline{\beta}_{r^*}))$ are assumed to be differentiable with respect to the elements of $h_{r^*}(x; \underline{\beta}_{r^*})$.

These mappings will allow us to substitute the parameters of the conditional distribution with compound functions of the model parameters in the estimation procedure described in Chapter 4, thus allowing us to estimate $\underline{\beta}_{r^*}$. In vector notation we write the mapping defined by Equation 3.3 as $\underline{\theta}_{r^*} = \underline{T}_{r^*}^{-1}(h_{r^*}(x; \underline{\beta}_{r^*}))$. This relationship between conditional distribution parameters and model parameters for a model may be used to define all the models previously discussed in this dissertation as well as those that arise in several other important applications. We shall see later in this chapter how many known models can be defined by the mappings given in Equation 3.3.

Our model includes a total of p manifest vector variables which are related to a single, independent variable that is measured with error. In the current chapter, we shall not include covariates measured without error, i.e., other variables related to the p manifest vectors which are not contaminated by measurement error. Error free covariates have previously been denoted by $\underline{Z}$. Note that adding covariates

measured without error may be an important extension of our model. In many applications, it may be necessary to control for additional error-free variables. A discussion of the possibility of future research to include such covariates, as well as an explanation of how they potentially could be handled in the model, are given in Chapter 7, where the summary discussion of this dissertation and future work are discussed.

The manifest variables shall, as mentioned at the beginning of this chapter, be denoted by the p_r -dimensional column vectors $\underline{Y}_r$, $r = 1, 2, \dots, p$, where the outcome of interest is one of the p vectors, i.e., $r^* \in \{1, 2, \dots, p\}$. Let x , the unobservable truth, be continuous with a known density denoted by $f(x | \underline{\theta}_x)$. The dimension of the column vector $\underline{\theta}_x$ is k_x .

For illustrative purposes, let us consider a generic application with three, one-dimensional manifest vector variables. One may wish to denote the outcome of interest, $\underline{Y}_{r^*}$, as Y_1 , say. Further, one may denote Y_2 as the measurement of the unobservable x (X in the notation defined in Chapter 2) which is usually considered to be an unbiased measurement, in which case $Y_2 = x + u$ with $E(u) = 0$, but not necessarily. The third observable variable, Y_3 , may denote the instrumental variable (W in the notation defined in Chapter 2) providing the identifying information. Recall from Chapter 2 that this is the variable providing additional information to identify the model. Parameters of the above model can be estimated using these two variables, Y_2 and Y_3 , in addition to the outcome of interest, Y_1 .

In other possible applications with more than three manifest variables, one may think of the extra $\underline{Y}_r$ as additional instrumental variables. We will show that the model parameters in the models defined for the p manifest vector variables, $\underline{Y}_1, \underline{Y}_2, \dots, \underline{Y}_p$, may be estimated utilizing their relations to x , whose mappings are described in vector notation as $\underline{T}_r(\underline{\theta}_r) = \underline{\gamma}_r = \underline{h}_r(x; \underline{\beta}_r)$, $r = 1, 2, \dots, p$.

A key assumption to the methodology that follows is conditional independence of the $\underline{Y}_r$ given x , $r = 1, 2, \dots, p$. Manifest variables which are not conditionally independent are grouped into manifest vectors that, together, satisfy the conditional independence assumption given x , when compared to the remaining manifest variables. This enables the GSME model to handle violations of the conditional independence assumption and to make full use of the information provided in the nondiagonal, variance-covariance matrix of those variables. We discuss this further in Chapter 4.

Since every manifest variable may be defined using the mappings similar to those for $\underline{Y}_{r^*}$ in Equation 3.3, we may further let, $\underline{\beta}_r$ be the l_r distinct parameters that describe the relation between $\underline{Y}_r$ and x , for each r . Also, since each $\underline{Y}_r$ has a known conditional distribution, each variable has k_r distributional parameters denoted by a vector $\underline{\theta}_r$. The relationship between these $\underline{\theta}_r$ and the model parameters, when solved for the distributional parameters, may be written in vector notation as

$$\underline{\theta}_r = \underline{T}_r^{-1} \left(\underline{h}_r(x; \underline{\beta}_r) \right), \quad r = 1, 2, \dots, p. \quad (3.4)$$

Assume that the mappings $\underline{T}_r^{-1} \left(\underline{h}_r(x; \underline{\beta}_r) \right)$ and $\underline{h}_r(x; \underline{\beta}_r)$ have the same properties as those stated above for $\underline{Y}_{r^*}$.

In the next chapter, we will see why it is convenient to be able to write the relation between all distributional parameters and all the model parameters as one vector. Concatenation of all of the column vectors of mappings, i.e., Equation 3.4 for all r , for the conditional distributional parameters of all the manifest variables and for those of x shall be written in one $(k_1 + k_2 + \dots + k_p + k_x) \times 1$ vector as

$$\underline{\theta} = \underline{T}^{-1} \left(\underline{h}(x; \underline{\beta}) \right), \quad (3.5)$$

where the $(l_1 + l_2 + \dots + l_p + k_x) \times 1$ vector of model parameters is denoted by

$$\underline{\beta} = (\underline{\beta}'_1, \underline{\beta}'_2, \dots, \underline{\beta}'_p, \underline{\theta}'_x)'$$

This vector of parameters, $\underline{\beta}$, shall be referred to as the “model parameters” since these are the parameters we intend to *estimate*, not to be confused with the vectors $\underline{\gamma}_r$ whose elements are the parameters we are interested in *modeling*, for each $\underline{Y}_r$. Further, the last k_x elements in $\underline{\theta}$ are the distributional parameters for x , which require no mappings, and therefore are simply equal to the last k_x elements in $\underline{\beta}$, which are of course $\underline{\theta}_x$. The mappings of $\underline{T}^{-1}(\underline{h}(x; \underline{\beta}))$ and $\underline{h}(x; \underline{\beta})$ will be written in short hand notation as $\underline{T}^{-1}$ and $\underline{h}$, respectively, whenever possible. We may also similarly abbreviate the mappings $\underline{T}_r^{-1}(\underline{h}_r(x; \underline{\beta}_r))$ and $\underline{h}_r(x; \underline{\beta}_r)$, respectively, as $\underline{T}_r^{-1}$ and $\underline{h}_r$. Since the column vector $\underline{T}^{-1}$ is simply the concatenation of the $\underline{T}_r^{-1}$, for all r , the entire concatenated vector, then, has the same properties as those stated for the mappings of the individual manifest variables. The same is obviously also true for $\underline{h}$.

This model shall be called the *generalized simple measurement error* (henceforth, GSME) *model*; simple because there is one covariate measured with error. By definition of a ME model, the GSME model therefore requires $p \geq 2$. This is to allow for an outcome of interest which is related to x and a measurement of x which is contaminated with error. The identifying information for our GSME model comes through the addition of at least one instrumental variable. In Chapter 5 we will show how our motivating example based on problems faced by the RERF on estimated radiation dose, discussed in Chapter 1, fits into the GSME model framework. In the next section of this chapter, we show by example how many well studied ME models can be defined using a GSME model.

3.2 Applications of the GSME Model

In this section, we will show how examples previously discussed in this dissertation, and some others of interest, fit into the GSME model setup. The examples considered describe scenarios for $\underline{Y}_r$, the outcome of interest, however, they are possible examples for any of the $\underline{Y}_r$. A strength of the GSME model

is the numerous and varied possibilities for its application. One GSME model could contain any combination of the examples presented below for its p manifest variables.

3.2.1 Exponential Family General Linear Models

One way a variable in the general exponential family may be modeled is through the well known Generalized Linear Model. Consider a GLM for $p_{r^*} = 1$ and Y_{r^*} in the single parameter, exponential family. We assume that $f(y_{r^*} | x; \underline{\theta}_{r^*})$ is known, where $\underline{\theta}_{r^*} = (\theta_{r^*}, \phi_{r^*})'$, θ_{r^*} is the canonical parameter, and ϕ_{r^*} is the dispersion parameter which, here, is assumed to be known. As we have seen in Subsection 2.4.3 of Chapter 2,

$$f(y_{r^*} | x; \underline{\theta}_{r^*}) = \exp[\{y_{r^*}\theta_{r^*} - b_{r^*}(\theta_{r^*})\}/a_{r^*}(\phi_{r^*}) + c_{r^*}(y_{r^*}, \phi_{r^*})].$$

In this case, we know that we are modeling $E(y_{r^*} | x; \underline{\theta}_{r^*}) = \mu_{r^*}(\eta_{r^*})$, where $\eta_{r^*} = \beta_{r^*0} + x\beta_{r^*1}$ and $\mu_{r^*}(\eta_{r^*})$ is specified through the choice of the link function. The parameters from Equation 3.1, in terms of the elements of $\underline{h}_{r^*}$, which we are interested in modeling, are

$$\begin{aligned}\gamma_{r^*1} &= h_{r^*1}(\eta_{r^*}) = E(Y_{r^*} | x) = \mu_{r^*}(\eta_{r^*}) \\ \gamma_{r^*2} &= h_{r^*2}(x; \beta_{r^*2}) = \beta_{r^*2},\end{aligned}$$

where β_{r^*2} is a constant which does not depend on x . Therefore the vector of model parameters is $\underline{\beta}_{r^*} = (\beta_{r^*0}, \beta_{r^*1}, \beta_{r^*2})'$.

The corresponding $\underline{T}_{r^*}(\underline{\theta}_{r^*})$ mappings from Equation 3.2 are

$$\begin{aligned}\gamma_{r^*1} &= T_{r^*1}(\underline{\theta}_{r^*}) = b'_{r^*}(\theta_{r^*}) \\ \gamma_{r^*2} &= T_{r^*2}(\underline{\theta}_{r^*}) = \phi_{r^*}\end{aligned}$$

Therefore, $\theta_{r^*} = b'^{-1}_{r^*}(\mu_{r^*}(\eta_{r^*}))$ and this is the relation between the conditional distribution parameter, θ_{r^*} , and the model parameters β_{r^*0} and β_{r^*1} , that will be

utilized in the estimation procedure which is discussed in Chapter 4. Further, we have $\phi_{r^*} = \beta_{r^*2}$. Thus we have established the mappings which were defined in Equation 3.3.

Note that there are no restrictions on the nonlinear mean function for variables in the general exponential family. For further details on the exponential family modeled via GLM's, see Subsection 2.4.3, where the model was first introduced, or, for example, McCullagh and Nelder [40].

3.2.2 Multinomial Model

There are many ME modeling situations, e.g. logistic and probit regression models, which require application to categorical outcomes. Therefore the ability to be applied to categorical data is of paramount importance for the GSME model. The outcome of interest in this example is categorical, having C_{r^*} categories. Here Y_{r^*} has known conditional cell probabilities given by $P(Y_{r^*} = j_{r^*} \mid x, \underline{\theta}_{r^*}) = \theta_{r^*s}$, $j_{r^*} = 1, 2, \dots, C_{r^*}$, $s = 1, 2, \dots, k_{r^*} = C_{r^*} - 1$, and we will write the θ_{r^*s} as π_{r^*s} since they represent conditional "probabilities". There are only required to be a total of $k_{r^*} = C_{r^*} - 1$ distributional parameters since $\pi_{r^*C_{r^*}} = 1 - \sum_{s=1}^{k_{r^*}} \pi_{r^*s}$.

Modeling the Means

If we are interested in modeling the mean, in Equation 3.1 we have

$$\begin{bmatrix} \gamma_{r^*1} \\ \gamma_{r^*2} \\ \vdots \\ \gamma_{r^*k_{r^*}} \end{bmatrix} = \begin{bmatrix} h_{r^*1}(x; \underline{\beta}_{r^*1}) \\ h_{r^*2}(x; \underline{\beta}_{r^*2}) \\ \vdots \\ h_{r^*k_{r^*}}(x; \underline{\beta}_{r^*k_{r^*}}) \end{bmatrix} = \begin{bmatrix} \mu_{r^*1}(x; \underline{\beta}_{r^*1}) \\ \mu_{r^*2}(x; \underline{\beta}_{r^*2}) \\ \vdots \\ \mu_{r^*k_{r^*}}(x; \underline{\beta}_{r^*k_{r^*}}) \end{bmatrix},$$

where one must specify the functional form of the $\mu_{r^*s}(\underline{x}; \underline{\beta}_{r^*s})$ functions, $s = 1, 2, \dots, k_{r^*}$. Further, for the relations defined in Equation 3.2 we have that

$$\begin{bmatrix} \gamma_{r^*1} \\ \gamma_{r^*2} \\ \vdots \\ \gamma_{r^*k_{r^*}} \end{bmatrix} = \begin{bmatrix} T_{r^*1}(\underline{\theta}_{r^*}) \\ T_{r^*2}(\underline{\theta}_{r^*}) \\ \vdots \\ T_{r^*k_{r^*}}(\underline{\theta}_{r^*}) \end{bmatrix} = \begin{bmatrix} \pi_{r^*1} \\ \pi_{r^*2} \\ \vdots \\ \pi_{r^*k_{r^*}} \end{bmatrix}.$$

Therefore the $\underline{T}_{r^*}^{-1}$ mapping in Equation 3.3 is the identity mapping. So, for example, if we modeled the mean as a logistic function we would have the conditional distributional parameters written in terms of the model parameters as

$$\pi_{r^*s} = \mu_{r^*s}(\underline{x}; \underline{\beta}_{r^*s}) = \frac{\exp(\beta_{r^*0s} + x\beta_{r^*1s})}{1 + \exp(\beta_{r^*0s} + x\beta_{r^*1s})},$$

$s = 1, 2, \dots, k_{r^*}$, where $\underline{\beta}_{r^*s} = (\beta_{r^*0s}, \beta_{r^*1s})'$ and β_{r^*0s} and β_{r^*1s} are the intercept and slope parameters, respectively, relating the conditional mean of $\underline{Y}_{r^*}$ to x for the s th conditional cell probability.

Modeling the Generalized Odds

Another example when the primary outcome of interest is a multinomial response could be to model the generalized odds as the functions of interest. In this case, we have the functions in Equation 3.1

$$\begin{bmatrix} \gamma_{r^*1} \\ \gamma_{r^*2} \\ \vdots \\ \gamma_{r^*k_{r^*}} \end{bmatrix} = \begin{bmatrix} h_{r^*1}(x; \underline{\beta}_{r^*1}) \\ h_{r^*2}(x; \underline{\beta}_{r^*2}) \\ \vdots \\ h_{r^*k_{r^*}}(x; \underline{\beta}_{r^*k_{r^*}}) \end{bmatrix} = \begin{bmatrix} GO_{r^*1}(x; \underline{\beta}_{r^*1}) \\ GO_{r^*2}(x; \underline{\beta}_{r^*2}) \\ \vdots \\ GO_{r^*k_{r^*}}(x; \underline{\beta}_{r^*k_{r^*}}) \end{bmatrix},$$

where $GO_{r^*s}(x; \underline{\beta}_{r^*s})$ is defined to be the probability of category s divided by the probability of the reference category, i.e., category C_{r^*} , say, where the probabilities are written in terms of x and $\underline{\beta}_{r^*s}$, e.g., a logistic function of the linear predictor.

The mappings of the conditional distributional parameters, which in this case are conditional cell probabilities, for use in Equation 3.2, are:

$$\begin{bmatrix} \gamma_{r^*1} \\ \gamma_{r^*2} \\ \vdots \\ \gamma_{r^*k_{r^*}} \end{bmatrix} = \begin{bmatrix} T_{r^*1}(\underline{\theta}_{r^*}) \\ T_{r^*2}(\underline{\theta}_{r^*}) \\ \vdots \\ T_{r^*k_{r^*}}(\underline{\theta}_{r^*}) \end{bmatrix} = \begin{bmatrix} \pi_{r^*1}/\pi_{r^*C_{r^*}} \\ \pi_{r^*2}/\pi_{r^*C_{r^*}} \\ \vdots \\ \pi_{r^*k_{r^*}}/\pi_{r^*C_{r^*}} \end{bmatrix},$$

where recall that $k_{r^*} = C_{r^*} - 1$. Solving for the parameters of the distribution in terms of x and $\underline{\beta}_{r^*}$ to define the mappings in Equation 3.3 results in

$$\begin{aligned} \pi_{r^*1} &= \pi_{r^*C_{r^*}} GO_{r^*1}(x; \underline{\beta}_{r^*1}) \\ \pi_{r^*2} &= \pi_{r^*C_{r^*}} GO_{r^*2}(x; \underline{\beta}_{r^*2}) \\ &\vdots \\ \pi_{r^*k_{r^*}} &= \pi_{r^*C_{r^*}} GO_{r^*k_{r^*}}(x; \underline{\beta}_{r^*k_{r^*}}) \\ \pi_{r^*C_{r^*}} &= \pi_{r^*C_{r^*}} GO_{r^*C_{r^*}}(x; \underline{\beta}_{r^*C_{r^*}}), \end{aligned}$$

where $GO_{r^*C_{r^*}}(x; \underline{\beta}_{r^*C_{r^*}}) = 1$. Since we know that the sum over the left hand side of the previous equation is one, we have that $\pi_{r^*C_{r^*}} + \pi_{r^*C_{r^*}} \sum_{s=1}^{k_{r^*}} GO_{r^*s}(x; \underline{\beta}_{r^*s}) = 1$. This implies that

$$\pi_{r^*C_{r^*}} = \frac{1}{1 + \sum_{s=1}^{k_{r^*}} GO_{r^*s}(x; \underline{\beta}_{r^*s})}$$

and that, for all s ,

$$\pi_{r^*s} = \frac{GO_{r^*s}(x; \underline{\beta}_{r^*s})}{1 + \sum_{s=1}^{k_{r^*}} GO_{r^*s}(x; \underline{\beta}_{r^*s})},$$

which defined the $\underline{T}_{r^*}^{-1}$ mappings in Equation 3.3.

3.2.3 Survival Time

In survival analysis problems, also known as the analysis of failure time data [33], the outcome, Y_{r^*} , is the time until an “event” of interest occurs and we model $Y_{r^*} = T = \text{survival time}$. It is standard to let T represent the outcome

in survival analysis problems, however, for consistency of this dissertation, we retain our notation, $Y_{r\bullet}$. As mentioned in Chapter 2, when we first introduced the concept of regression calibration, Prentice [44] modeled the hazard function in a survival analysis, measurement error problem. Therefore, we know survival analysis problems are an important application of our GSME model.

For the purposes of this example, suppose that $Y_{r\bullet}$ is the time until the diagnosis of leukemia for Atomic-bomb survivors (see Chapter 1). The model of interest is the hazard function of $Y_{r\bullet}$ as a function of x , true radiation dose. A very general definition of a hazard function is a function which describes the chances that an individual in the study observes an “event” in the next instant of time (see, for example, [35, 33]). A complication arises in this example due to the fact that an “event”, or diagnosis of leukemia, may be the result of background causes or of radiation dose, x . In other words, there will be a background rate of occurrences that is independent of exposure, x . To incorporate this reality into our modeling framework, we write the hazard function as the sum of two different hazards; one a function of the unknown x and the other a background hazard independent of x . We show, in detail, in Chapter 5 how this motivating example can be set up to fit into the GSME model framework.

So, one drawback of the GSME model is that it is not generally applicable to all survival analysis problems. The survival analysis problems that may fit into the framework of the GSME model, would most likely need to be modified or adjusted in some manner in order to do so. It would seem that survival analysis problems would need to be handled on a case by case basis, with special care in making sure to meet the requirements of the GSME model.

3.2.4 Multivariate Regression Model

For multivariate nonlinear regression we model a p_{r^*} -dimensional, multivariate response, $\underline{Y}_{r^*}$. The model is:

$$\begin{bmatrix} Y_{r^*1} \\ Y_{r^*2} \\ \vdots \\ Y_{r^*p_{r^*}} \end{bmatrix} = \begin{bmatrix} \mu_{r^*1}(x; \underline{\beta}_{r^*1}) \\ \mu_{r^*2}(x; \underline{\beta}_{r^*2}) \\ \vdots \\ \mu_{r^*p_{r^*}}(x; \underline{\beta}_{r^*p_{r^*}}) \end{bmatrix} + \underline{e}_{r^*},$$

where $\mu_{r^*s}(\cdot)$ is an arbitrary nonlinear mean function for the s th variable, $s = 1, 2, \dots, p_{r^*}$, and we assume $E(\underline{e}_{r^*}) = \underline{0}$ and $Var(\underline{e}_{r^*}) = \Sigma_{e_{r^*}}$. Conditional dependence in the multivariate regression model means that $\Sigma_{e_{r^*}}$ is not a diagonal matrix. The elements on the diagonal are the variances of the Y_{r^*s} , $s = 1, 2, \dots, p_{r^*}$, and the ij th element is the covariance between Y_{r^*i} and Y_{r^*j} , for $i \neq j$.

In multivariate linear regression models, the $\mu_{r^*s}(\cdot)$ are linear functions. The linear model in matrix form is

$$\underline{Y}_{r^*} = \mathbf{B}_{r^*} \underline{X} + \underline{e}_{r^*},$$

where $\underline{Y}_{r^*}$ is a $p_{r^*} \times 1$ vector, $\mathbf{B}_{r^*}$ is a $p_{r^*} \times 2$ matrix, and $\underline{X} = (1, x)'$ is a 2×1 vector. The s th equation would be $Y_{r^*s} = \beta_{r^*0s} + x\beta_{r^*1s} + e_{r^*s}$, $s = 1, 2, \dots, p_{r^*}$. Assume that $f_{r^*}(\underline{y}_{r^*} | x; \underline{\theta}_{r^*})$ is known and recall that the dimension of $\underline{\theta}_{r^*}$ is $k_{r^*} \times 1$ and it contains the parameters $\underline{\mu}_{r^*}$ (i.e., the means) and the distinct elements of the variance-covariance matrix $\Sigma_{e_{r^*}}$, say σ_{ab} , for all $a = 1, 2, \dots, p_{r^*}$, $b = 1, 2, \dots, p_{r^*}$, and $a \leq b$, where σ_{ab} is a variance when $a = b$, and a covariance when not. For the linear model, the functions of x and $\underline{\beta}_{r^*}$ that we wish to model, in terms of the functions for Equation 3.1, are

$$\begin{bmatrix} \gamma_{r^*1} \\ \gamma_{r^*2} \\ \vdots \\ \gamma_{r^*p_{r^*}} \\ \gamma_{r^*p_{r^*}+1} \\ \vdots \\ \gamma_{r^*k_{r^*}} \end{bmatrix} = \begin{bmatrix} h_{r^*1}(x; \underline{\beta}_{r^*1}) \\ h_{r^*2}(x; \underline{\beta}_{r^*2}) \\ \vdots \\ h_{r^*p_{r^*}}(x; \underline{\beta}_{r^*p_{r^*}}) \\ h_{r^*p_{r^*}+1}(x; \underline{\beta}_{r^*p_{r^*}+1}) \\ \vdots \\ h_{r^*k_{r^*}}(x; \underline{\beta}_{r^*k_{r^*}}) \end{bmatrix} = \begin{bmatrix} \beta_{r^*01} + x\beta_{r^*11} \\ \beta_{r^*02} + x\beta_{r^*12} \\ \vdots \\ \beta_{r^*0p_{r^*}} + x\beta_{r^*1p_{r^*}} \\ \beta_{r^*p_{r^*}+1} \\ \vdots \\ \beta_{r^*k_{r^*}} \end{bmatrix},$$

where the $\beta_{r^*p_{r^*}+1}, \beta_{r^*p_{r^*}+2}, \dots, \beta_{r^*k_{r^*}}$ are constants and k_{r^*} is a number such to allow the dimension of the sub-vector $(\gamma_{r^*p_{r^*}+1} \gamma_{r^*p_{r^*}+2} \dots, \gamma_{r^*k_{r^*}})'$ to be equal the number of distinct variances and covariances σ_{ab} , $a = 1, 2, \dots, p_{r^*}$, $b = 1, 2, \dots, p_{r^*}$, $a \leq b$. The corresponding mappings of the distributional parameters in Equation 3.2 are

$$\begin{bmatrix} \gamma_{r^*1} \\ \gamma_{r^*2} \\ \vdots \\ \gamma_{r^*p_{r^*}} \\ \gamma_{r^*p_{r^*}+1} \\ \vdots \\ \gamma_{r^*k_{r^*}} \end{bmatrix} = \begin{bmatrix} T_{r^*1}(\underline{\theta}_{r^*}) \\ T_{r^*2}(\underline{\theta}_{r^*}) \\ \vdots \\ T_{r^*p_{r^*}}(\underline{\theta}_{r^*}) \\ T_{r^*p_{r^*}+1}(\underline{\theta}_{r^*}) \\ \vdots \\ T_{r^*k_{r^*}}(\underline{\theta}_{r^*}) \end{bmatrix} = \begin{bmatrix} \mu_{r^*1} \\ \mu_{r^*2} \\ \vdots \\ \mu_{r^*p_{r^*}} \\ \sigma_{r^*p_{r^*}+1} \\ \vdots \\ \sigma_{r^*k_{r^*}} \end{bmatrix}.$$

In writing these mappings of the conditional distribution parameters $\underline{\theta}_{r^*}$, it is easily seen why the dimension of the sub-vector of γ_{r^*} 's, $(\gamma_{r^*p_{r^*}+1} \gamma_{r^*p_{r^*}+2} \dots, \gamma_{r^*k_{r^*}})'$, must contain enough elements to allow all the distinct elements of the variance-covariance matrix $\Sigma_{e_{r^*}}$ to be mapped, along with the means, into the parameter space Γ_{r^*} .

One important attribute of this application of the GSME model is that performing a multivariate regression (linear or nonlinear) may be a potential solution

to the problem which can arise concerning the assumption of conditional independence of the manifest variables. If the assumption of conditional independence is violated or in serious question, then grouping those variables that are suspected to be conditionally dependent together and defining them as one manifest vector variable, $\underline{Y}_r$, that requires a multivariate regression model, makes use of the extra information contained in the covariances, without violating of the key assumption of conditional independence.

3.2.5 Nonlinear Model

The full generality of the GSME model allows for models with nonlinear mean functions which are not necessarily functions of a linear relationship between x and the parameters which describe the relation between x and the outcome of interest. This is an improvement over generalized linear ME models, discussed, for example, in Carroll et al. [13], whose mean functions are in terms of a strictly linear relation between x and the model parameters. For the outcome of interest, the nonlinear model is given by

$$Y_{r\bullet} = h_{r\bullet}(x; \underline{\beta}_{r\bullet}) + e_{r\bullet},$$

where $h_{r\bullet}(\cdot)$ is an arbitrary nonlinear mean function and $E(e_{r\bullet}) = 0$. This model was previously discussed in Section 2.4 of Chapter 2. It is conventional to define a ME model as nonlinear when $\underline{\beta}_{r\bullet}$ enters the mean function in a nonlinear manner or when the mean function is nonlinear in x . Nonlinear models describe a large class of models, which may be applied to both continuous and discrete outcomes. The focus of Section 2.4 was initially on continuous outcomes under the assumption of normality, before moving on to outcomes in the exponential family modeled through GLM's.

Let us first consider a simple nonlinear model, assuming normality, given as

$$Y_{r^*} = \beta_{r^*0} + x\beta_{r^*1} + x^2\beta_{r^*2} + e_{r^*},$$

where $E(e_{r^*}) = E(xe_{r^*}) = E(x^2e_{r^*}) = 0$. This model was discussed in Section 2.4, where it was shown that a bias in the naive estimators for nonlinear models exists, and that the bias in the estimator for the β_{r^*2} parameter is even greater than that for the estimator of β_{r^*1} . Thus, the ability of the GSME model to be applied to this scenario, and reduce the problems of the bias through its methods, is of notable importance. Assuming a constant variance for the conditional normal distribution of Y_{r^*} , the parameters which are modeled in terms of the mappings given by Equation 3.1 are

$$\begin{bmatrix} \gamma_{r^*1} \\ \gamma_{r^*2} \end{bmatrix} = \begin{bmatrix} h_{r^*1}(x; \underline{\beta}_{r^*1}) \\ h_{r^*2}(x; \underline{\beta}_{r^*2}) \end{bmatrix} = \begin{bmatrix} \beta_{r^*0} + x\beta_{r^*1} + x^2\beta_{r^*2} \\ \beta_{r^*3} \end{bmatrix},$$

where $\underline{\beta}_{r^*1} = (\beta_{r^*0}, \beta_{r^*1}, \beta_{r^*2})'$ and $\underline{\beta}_{r^*2}$ denotes a 1×1 vector consisting of the scalar β_{r^*3} , which is a constant. Under the current assumptions, the distribution of Y_{r^*} given x is normal with conditional parameters, the mean, μ_{r^*} , and the constant variance, $\sigma_{r^*r^*}$. Therefore, $\underline{\theta}_{r^*} = (\mu_{r^*}, \sigma_{r^*r^*})'$. The mappings defined in Equation 3.2, which correspond to those given above are

$$\begin{bmatrix} \gamma_{r^*1} \\ \gamma_{r^*2} \end{bmatrix} = \begin{bmatrix} T_{r^*1}(\underline{\theta}_{r^*}) \\ T_{r^*2}(\underline{\theta}_{r^*}) \end{bmatrix} = \begin{bmatrix} \mu_{r^*} \\ \sigma_{r^*r^*} \end{bmatrix}.$$

As alluded to in Section 3.1 of this chapter, this permits us to write the conditional distribution, given here by:

$$f(y_{r^*} \mid x; \underline{\theta}_{r^*}) = \frac{1}{\sqrt{2\pi\sigma_{r^*r^*}}} \exp \left\{ -\frac{(y_{r^*} - \mu_{r^*})^2}{2\sigma_{r^*r^*}} \right\},$$

in terms of the model parameters, i.e., those we ultimately intend to estimate, as:

$$f(y_{r^*} \mid x; \underline{\beta}_{r^*}) = \frac{1}{\sqrt{2\pi}\beta_{r^*3}} \exp \left\{ -\frac{(y_{r^*} - (\beta_{r^*0} + x\beta_{r^*1} + x^2\beta_{r^*2}))^2}{2\beta_{r^*3}} \right\},$$

where $\underline{\beta}_{r^*}$, which equals $(\beta_{r^*0}, \beta_{r^*1}, \beta_{r^*2}, \beta_{r^*3})'$, written on the left hand side of the preceding equation should be understood to be a symbolic representation of the mappings $\underline{T}_{r^*}^{-1}(\underline{h}_{r^*}(x; \underline{\beta}_{r^*}))$. Since the density is a function of $\underline{\beta}_{r^*}$, this will allow us to estimate the model parameters, and the process of how this estimation is performed, is shown in the next chapter.

Let us now look at how these mappings are defined for another common density, where it will not be necessary to specify a model to do so. Assume that Y_{r^*} follows a Gamma distribution. One common way to model a Gamma distribution is through a GLM, however, the mean of a Gamma distribution is not required to be a nonlinear function of a linear predictor. The mean could be modeled by some arbitrary nonlinear function, which is the reason for including this example in the current subsection on nonlinear models. The “standard form” of the density for a Gamma distribution, with parameters α and λ , i.e., $\text{Gamma}(\alpha, \lambda)$, is

$$f(y_{r^*} \mid x; \underline{\theta}_{r^*}) = \frac{(y_{r^*})^{\alpha-1} (\lambda)^\alpha \exp(-\lambda y_{r^*})}{\Gamma(\alpha)},$$

where $\underline{\theta}_{r^*} = (\alpha, \lambda)$ and for notational convenience, in this example only, we drop the subscript of r^* on the specific distributional parameters. We retain this notation only for the random variable itself, y_{r^*} , and the vector denoting its conditional distribution parameters, $\underline{\theta}_{r^*}$. In this “standard form”, we have that $E(y_{r^*}) = \alpha/\lambda$ and $\text{Var}(y_{r^*}) = \alpha/\lambda^2$. In order to show how the above model fits into the GSME model framework, we reparameterize the conditional distribution so that it is in the form for the exponential family. The reparameterization $\mu = \nu/\lambda$ and $\nu = \alpha$ results in $E(y_{r^*}) = \mu$ and $\text{Var}(y_{r^*}) = \mu^2/\nu$. Often one may not work with the Gamma density in its “standard form” given above. If

the preference is to write the density in the form of the exponential family, the previous reparameterization would not be needed and one could directly proceed with defining the mappings for $\underline{h}_{r^\bullet}$ and $\underline{T}_{r^\bullet}$, for the density in the preferred form. Continuing in the present context, the density function becomes

$$\begin{aligned} f(y_{r^\bullet} \mid x; \underline{\theta}_{r^\bullet}^\#) &= \frac{(y_{r^\bullet})^{\nu-1} (\nu/\mu)^\nu \exp(-(\nu/\mu)y_{r^\bullet})}{\Gamma(\nu)} \\ &= \exp \left[\frac{\{y_{r^\bullet}(-1/\mu) + \log(-1/\mu)\}}{1/\nu} \right. \\ &\quad \left. + \nu \log \nu + (\nu - 1) \log y_{r^\bullet} - \log \Gamma(\nu) \right], \end{aligned}$$

where $\underline{\theta}_{r^\bullet}^\#$ represents the conditional distribution parameters of the density in this form (the superscript “#” indicates a different set of conditional parameters than those in $\underline{\theta}_{r^\bullet}$ for the “standard” density) with elements which are the canonical parameter, $\theta = -1/\mu$, and $\phi = 1/\nu$. In the notation for the exponential family we have that $a(\phi) = \phi$, $b(\theta) = -\log(-\theta)$, and $c(y_{r^\bullet}, \phi) = (1/\phi) \log(1/\phi) + ((1/\phi) - 1) \log y_{r^\bullet} - \log \Gamma(1/\phi)$. Further, we know that $E(y_{r^\bullet}) = b'(\theta) = -1/\theta = \mu$ and $Var(y_{r^\bullet}) = a(\phi)b''(\theta) = \phi(1/\theta^2) = (1/\nu)\mu^2$. In this parameterization, we have as our $\underline{h}_{r^\bullet}(x; \underline{\beta}_{r^\bullet})$ functions from Equation 3.1:

$$\begin{bmatrix} \gamma_{r^\bullet 1} \\ \gamma_{r^\bullet 2} \end{bmatrix} = \begin{bmatrix} h_{r^\bullet 1}(x; \underline{\beta}_{r^\bullet 1}) \\ h_{r^\bullet 2}(x; \underline{\beta}_{r^\bullet 2}) \end{bmatrix} = \begin{bmatrix} \mu(x; \underline{\beta}_{r^\bullet 1}) \\ \nu \end{bmatrix},$$

where $\nu = \beta_{r^\bullet 2}$ is a constant scalar and $\mu(\cdot)$ is the nonlinear mean function written in terms of x and $\underline{\beta}_{r^\bullet 1}$, which, again, is not required to be written as a function of a linear predictor. In practice, one specifies a form for the mean function. For the $\underline{T}_{r^\bullet}$ mappings defined in Equation 3.2, we have

$$\begin{bmatrix} \gamma_{r^\bullet 1} \\ \gamma_{r^\bullet 2} \end{bmatrix} = \begin{bmatrix} T_{r^\bullet 1}(\underline{\theta}_{r^\bullet}^\#) \\ T_{r^\bullet 2}(\underline{\theta}_{r^\bullet}^\#) \end{bmatrix} = \begin{bmatrix} b'(\theta) \\ 1/\phi \end{bmatrix},$$

where, again, $b'(\theta) = -1/\theta$. Using these mappings we can solve for the elements of $\underline{\theta}_{r*}^\# = (\theta, \phi)$, giving $\theta = -1/\mu(x; \underline{\beta}_{r*1})$ and $\phi = 1/\nu$, which is a constant. Again, if it is not desired to use the “standard” density for a Gamma distribution, the definition of the model mappings stops here. Otherwise if there is a preference to work with its “standard” density, we reparameterize back to the parameters of the “standard form” of the density. In terms of the parameters from the original density, $\underline{\theta}_{r*} = (\alpha, \lambda)$, for Equation 3.3, i.e., the elements of the mapping $\underline{\theta}_{r*} = T_{r*}^{-1} \left(h_{r*}(x; \underline{\beta}_{r*1}) \right)$, we have $\lambda = \alpha/\mu(x; \underline{\beta}_{r*1})$ and $\alpha = \beta_{r*2}$, a constant.

3.3 Model Identification

As first mentioned in Chapter 2, an important question that arises with the study of ME models is the identification of a model’s parameters. Again, if the parameters of a model are identified, we may also say the model itself is identified. To consider the identification of the GSME model, we will need to make reference to the joint marginal distribution of the manifest variables. Due to the assumption of conditional independence, this joint distribution is given by

$$f(\underline{y}_1, \underline{y}_2, \dots, \underline{y}_p) = \int_x \prod_{r=1}^p f_r(\underline{y}_r | x; \underline{\theta}_r) f(x | \underline{\theta}_x) dx,$$

where $\int_x$ denotes integration over x and $f_r(\underline{y}_r | x; \underline{\theta}_r)$ is a generic representation for the distribution function of the r th manifest variable, where $\underline{y}_r$ may be a combination of variables which are grouped together due to a violation of the conditional independence assumption. Recall that $\underline{\theta}_r$ are the conditional distribution parameters for the r th manifest variable. This joint marginal distribution, the variables which make it up, and the conditional distributions of the $\underline{Y}_r$ are discussed in detail in Chapter 4. Further, we may write the joint marginal distribution as $f(\underline{y}_1, \underline{y}_2, \dots, \underline{y}_p) = f_Y(\underline{y}; \underline{\omega})$, where $(\underline{y}'_1, \underline{y}'_2, \dots, \underline{y}'_p)' = (\underline{y})$ and $\underline{\omega}$ is the vector of parameters of the joint distribution that lie in a parameter space Ω . It is seen that these joint distributional parameters must be in terms of the parameters

$\underline{\theta}_x$ and $\underline{\theta}_r$, for all r . Further, due to the mappings defined by Equation 3.5, i.e., $\underline{\theta} = \underline{T}^{-1}(\underline{h}(x; \underline{\beta}))$, this is equivalent to saying that the elements of $\underline{\omega}$ must also be in terms of the model parameters, i.e., the elements of $\underline{\beta} = (\underline{\beta}'_1, \underline{\beta}'_2, \dots, \underline{\beta}'_p, \underline{\theta}'_x)'$. Since $\underline{\beta}$ is the vector of parameters that must ultimately be estimated, these are the parameters we must consider in discussing identification. The parameter space of $\underline{\beta}$ is denoted B . By the preceding discussion, there exists a mapping from B onto Ω (the image of B) such that every parameter vector $\underline{\beta}$ in B gets mapped onto an $\underline{\omega}$ in Ω , which defines the parameter space of $f(\underline{y}_1, \underline{y}_2, \dots, \underline{y}_p)$. Denote this mapping by $\underline{\omega}(\underline{\beta}) : B \rightarrow \Omega$.

Applying Fuller's [24] definition of identifiability, stated earlier in Chapter 2, to the GSME model would translate to the vector $\underline{\beta}$ being identified if, for any $\underline{\beta} \in B$ and $\underline{\beta}^* \in B$, $\underline{\beta} \neq \underline{\beta}^*$, there exists an arbitrary vector $\underline{a}$, with elements $(\underline{a}'_1, \underline{a}'_2, \dots, \underline{a}'_p)'$, such that $f_Y(\underline{a}; \underline{\omega}(\underline{\beta})) \neq f_Y(\underline{a}; \underline{\omega}(\underline{\beta}^*))$.

A necessary condition for $\underline{\beta}$ to be identified is $\dim(\underline{\omega}) \geq \dim(\underline{\beta})$, where $\dim(\cdot)$ is the dimension of the given vector. Under the present GSME model, it will often be impossible to determine $\dim(\underline{\omega})$ since $f(\underline{y}_1, \underline{y}_2, \dots, \underline{y}_p)$, the joint distribution of the manifest variables, is usually unknown and therefore $\underline{\omega}$ is unknown.

The following two theorems provide necessary and sufficient conditions under which the parameters of the GSME model are identified. First a lemma that is used in the proofs of these theorems.

LEMMA 1: If the parameters, $\underline{\omega}$, of the joint distribution $f_Y(\underline{y}; \underline{\omega})$ are identified, then the mapping, $\underline{\omega}(\underline{\beta})$, which maps the parameter space B onto the parameter space Ω , must be a function, i.e., each $\underline{\beta} \in B$ is mapped into one and only one vector of Ω , where $\underline{\beta} \in B$ and $\underline{\omega}(\underline{\beta}) \in \Omega$.

Proof: Assume not. Then there exists a $\underline{\beta}$ that is mapped into two different $\underline{\omega}$'s, say $\underline{\omega}_1$ and $\underline{\omega}_2$. Since $\underline{\omega}_1 = \underline{\omega}(\underline{\beta})$ and $\underline{\omega}_2 = \underline{\omega}(\underline{\beta})$, i.e., both are images of

the same $\underline{\beta}$, the densities of those vectors must be equal. In other words, we have $f_Y(\underline{a}; \underline{\omega}_1) = f_Y(\underline{a}; \underline{\omega}_2)$, for all $\underline{a}$. This contradicts the identifiability of $\underline{\omega}$. $\diamond$

(Note that we shall use, throughout, the symbol $\diamond$ to indicate the completion of the proof to a Theorem, Lemma, etc.)

THEOREM 1: Assume as a regularity condition that $\underline{\omega}$, the parameters of the joint distribution of the manifest variables, $f_Y(\underline{y}; \underline{\omega})$, are identified. Then, the vector of model parameters, $\underline{\beta}$, is identified if and only if the mapping of B onto Ω , denoted by $\underline{\omega}(\underline{\beta})$, invertible.

Proof: (sufficiency) Assume that $\underline{\omega}(\underline{\beta})$ is invertible. Therefore, $\underline{\omega}(\underline{\beta})$ must be one-to-one. Then for $\underline{\beta} \neq \underline{\beta}^*$, we know $\underline{\omega}(\underline{\beta}) \neq \underline{\omega}(\underline{\beta}^*)$. Also, by the identifiability of $\underline{\omega}$, we then have that $f_Y(\underline{a}; \underline{\omega}(\underline{\beta})) \neq f_Y(\underline{a}; \underline{\omega}(\underline{\beta}^*))$ for at least one $\underline{a}$ if $\underline{\omega}(\underline{\beta}) \neq \underline{\omega}(\underline{\beta}^*)$, and therefore $f_Y(\underline{a}; \underline{\beta}) \neq f_Y(\underline{a}; \underline{\beta}^*)$, which represent the densities written as functions of $\underline{\beta}$ and $\underline{\beta}^*$, for at least that same $\underline{a}$. Thus, $\underline{\beta}$ is identified.

(necessity) Assume $\underline{\beta}$ is identified. Then there exists an $\underline{a}$ such that $f(\underline{a}; \underline{\omega}(\underline{\beta})) \neq f(\underline{a}; \underline{\omega}(\underline{\beta}^*))$ for every $\underline{\beta}$ and $\underline{\beta}^*$ such that $\underline{\beta} \neq \underline{\beta}^*$. Now, by Lemma 1, $\underline{\omega}(\underline{\beta})$ is a function. Therefore we know that one $\underline{\beta}$ is not mapped to two different $\underline{\omega}$'s, thus we are only left with showing that for $\underline{\beta} \neq \underline{\beta}^*$ we have $\underline{\omega}(\underline{\beta}) \neq \underline{\omega}(\underline{\beta}^*)$. We do this by showing a contradiction.

Assume, now, that $\underline{\omega}(\underline{\beta})$ is not one-to-one. Then there exists a $\underline{\beta} \neq \underline{\beta}^*$ such that $\underline{\omega}(\underline{\beta}) = \underline{\omega}(\underline{\beta}^*)$, i.e., two different $\underline{\beta}$'s are mapped to the same $\underline{\omega}$. This contradicts the identifiability of $\underline{\omega}$, as we know there exists an $\underline{a}$ such that $f(\underline{a}; \underline{\omega}(\underline{\beta})) \neq f(\underline{a}; \underline{\omega}(\underline{\beta}^*))$, for all $\underline{\beta} \neq \underline{\beta}^*$. Therefore, the mapping $\underline{\omega}(\underline{\beta})$ is one-to-one. By definition given at the beginning of this section, we know the mapping $\underline{\omega}(\underline{\beta})$ is also onto. Thus $\underline{\omega}(\underline{\beta})$ is invertible and the proof is complete. $\diamond$

Prior to stating the next theorem, we give a definition. A *just identified* set of parameters, in this application, exist when $\dim(\underline{\omega}) = \dim(\underline{\beta})$ and $\underline{\omega}(\underline{\beta})$ is invertible. This definition is used in the next theorem.

THEOREM 2: Assume that $\underline{\omega}$, the vector of parameters of $f_Y(y; \underline{\omega})$, is identified. If $\dim(\underline{\omega}) > \dim(\underline{\beta})$, then the vector $\underline{\beta}$ is identified if there exists a just identified, proper subvector of $\underline{\omega}(\underline{\beta})$, denoted by $\underline{\omega}_1(\underline{\beta}) = \underline{\omega}_1$ (i.e., there exists a subvector of functions, $\underline{\omega}_1(\underline{\beta})$, having the same dimension as $\underline{\beta}$, such that $\underline{\omega}_1^{-1}(\cdot)$ exists satisfying $\underline{\omega}_1^{-1}\underline{\omega}_1(\underline{\beta}) = \underline{\beta}$ for all $\underline{\beta} \in B$).

Proof: Assume there is a just identified, proper subvector of $\underline{\omega}(\underline{\beta})$ denoted by $\underline{\omega}_1(\underline{\beta}) = \underline{\omega}_1$. So, $\underline{\omega}_1(\underline{\beta})$ is such that $\dim(\underline{\omega}_1(\underline{\beta})) = \dim(\underline{\beta}) < \dim(\underline{\omega})$ and it is invertible, by definition. This implies that $\underline{\beta} = \underline{\omega}_1^{-1}(\underline{\omega}_1(\underline{\beta}))$. Now let $\underline{\omega} = (\underline{\omega}'_1, \underline{\omega}'_2)'$ and define $\underline{\omega}^*(\underline{\omega}) = \underline{\omega}_1^{-1}(\underline{\omega}_1)$, where $\underline{\omega}_1^{-1}$ denotes the inverse of the invertible subvector $\underline{\omega}_1(\underline{\beta})$. Then $\underline{\omega}^*(\underline{\omega}(\underline{\beta})) = \underline{\omega}_1^{-1}(\underline{\omega}_1(\underline{\beta}))$ which is equal to $\underline{\beta}$. Thus $\underline{\omega}^*(\cdot)$ is the inverse of $\underline{\omega}(\cdot)$. By Theorem 1, $\underline{\beta}$ is identified. $\diamond$ [Note: In the existence of an invertible, proper subvector, $\underline{\omega}_1(\underline{\beta})$, of $\underline{\omega}(\underline{\beta})$, we have

$$\begin{bmatrix} \underline{\omega}_1 \\ \underline{\omega}_2 \end{bmatrix} = \begin{bmatrix} \underline{\omega}_1(\underline{\beta}) \\ \underline{\omega}_2(\underline{\beta}) \end{bmatrix} = \begin{bmatrix} \underline{\omega}_1(\underline{\beta}) \\ \underline{\omega}_2(\underline{\omega}_1^{-1}(\underline{\omega}_1)) \end{bmatrix} = \underline{\omega}(\underline{\beta}).$$

That is, the $\underline{\omega}_2$ parameters are redundant with $\underline{\omega}_1$.]

To see these theorems applied to a simple, easily seen case, consider the simple linear ME model with normal errors defined in Chapter 2. Using the notation from the current chapter the model would be

$$Y_1 = \beta_{10} + \beta_{11}x + e$$

$$Y_2 = x + u,$$

where $e \sim N(0, \sigma_e^2)$ and $u \sim N(0, \sigma_u^2)$. Further, under the structural model, assuming normality, we have $x \sim N(\mu_x, \sigma_x^2)$. Therefore $\underline{\beta} = (\beta_{10}, \beta_{11}, \sigma_e^2, \sigma_u^2, \mu_x, \sigma_x^2)'$ and $\dim(\underline{\beta}) = 6$. (Note that the notation for a variance has changed from that defined in Chapter 2. We shall use, for example, σ_x^2 in place of σ_{xx} , because attempting to use the double subscript notation for a variance, with the numerous subscripts

required for the GSME model, often proves to be too cumbersome and allows too easily for confusion. Therefore we will use the notation with the squared exponent to represent a variance for the remainder of this dissertation.)

The distribution of the joint density is

$$f(y_1, y_2) \sim MVN \left(\begin{bmatrix} \mu_{Y_1} & \mu_{Y_2} \end{bmatrix}', \begin{bmatrix} \sigma_{Y_1}^2 & \sigma_{Y_1 Y_2} \\ \sigma_{Y_1 Y_2} & \sigma_{Y_2}^2 \end{bmatrix} \right)$$

and $\dim(\underline{\omega}) = 5$. Recall that a necessary condition for $\underline{\beta}$ to be identified, stated prior to Lemma 1, was if $\dim(\underline{\omega}) \geq \dim(\underline{\beta})$. So the first indication that this model may not be identified without additional information, as we already know that it is not, is that $\dim(\underline{\beta}) > \dim(\underline{\omega})$. We know because of the dimensions, we may not apply Theorem 2. Further, we know that $\underline{\omega}(\underline{\beta})$ is not invertible, such to give $\underline{\omega}^{-1}(\underline{\omega}(\underline{\beta})) = \underline{\beta}$, so by Theorem 1 we may not conclude $\underline{\beta}$ is identified. If we can assume a known parameter in $\underline{\beta}$, then $\dim(\underline{\beta}) = \dim(\underline{\omega})$ and we then have hopes that $\underline{\beta}$ is identified. As we have seen in Subsection 2.2.2, if we assume σ_u^2 is known, then the solution, $\hat{\underline{\omega}}^{-1}(\hat{\underline{\omega}}(\hat{\underline{\beta}})) = \hat{\underline{\beta}}$, is

$$\begin{aligned} \hat{\beta}_{11} &= (m_{Y_2}^2 - \sigma_u^2)^{-1} m_{Y_2 Y_1}, \\ (\hat{\sigma}_x^2, \hat{\sigma}_e^2) &= (m_{Y_2}^2 - \sigma_u^2, m_{Y_1}^2 - \hat{\beta}_{11} m_{Y_2 Y_1}), \\ (\hat{\sigma}_{\omega_1}, \hat{\sigma}_{\omega_2}) &= (\sqrt{\sigma_e^2}, \sqrt{\sigma_x^2}) \quad (\hat{\sigma}_{\omega_3}, \hat{\sigma}_{\omega_4}) \end{aligned}$$

So, in this case where σ_u^2 is known, we may apply Theorem 1 stating that the vector of the parameters of the joint marginal distribution is invertible and thus $\underline{\beta}$ is identified.

Now, with the addition of an instrumental variable which is linearly related to x , we also have that $Y_3 = \beta_{30} + \beta_{31}x + v$, where we may assume that $v \sim N(0, \sigma_v^2)$. So now

$$\underline{\beta} = (\beta_{10}, \beta_{11}, \sigma_e^2, \sigma_u^2, \beta_{30}, \beta_{31}, \sigma_v^2, \mu_x, \sigma_x^2)'$$

with $\dim(\underline{\beta}) = 9$ and

$$f(y_1, y_2, y_3) \sim MVN \left([\mu_{Y_1}, \mu_{Y_2}, \mu_{Y_3}]', \begin{bmatrix} \sigma_{Y_1}^2 & \sigma_{Y_1 Y_2} & \sigma_{Y_1 Y_3} \\ \sigma_{Y_1 Y_2} & \sigma_{Y_2}^2 & \sigma_{Y_2 Y_3} \\ \sigma_{Y_1 Y_3} & \sigma_{Y_2 Y_3} & \sigma_{Y_3}^2 \end{bmatrix} \right).$$

So under this new joint marginal distribution, $\dim(\underline{\omega}) = 9$, and therefore, by the necessary condition stated prior to Lemma 1, we know we have hopes that the model may be identified, as we already know it is, since $\dim(\underline{\beta}) \leq \dim(\underline{\omega})$. Again, because of the dimensions of the parameter vectors, we cannot apply Theorem 2. We do know, however, that we can arrive at $\underline{\hat{\omega}}^{-1}(\underline{\hat{\omega}}(\underline{\hat{\beta}})) = \underline{\hat{\beta}}$, which is given by

$$\begin{aligned} \hat{\beta}_{11} &= \frac{m_{Y_1 Y_3}}{m_{Y_2 Y_3}} \\ (\hat{\sigma}_x^2, \hat{\sigma}_e^2) &= \left(\frac{m_{Y_2 Y_1} m_{Y_2 Y_3}}{m_{Y_1 Y_3}}, m_{Y_1}^2 - \hat{\beta}_{11} m_{Y_2 Y_1} \right) \\ (\hat{\mu}_x, \hat{\beta}_{10}) &= (\bar{Y}_2, \bar{Y}_1 - \hat{\beta}_{11} \bar{Y}_2) \\ (\hat{\beta}_{31}, \hat{\beta}_{30}) &= \left(\frac{m_{Y_1 Y_3}}{m_{Y_2 Y_1}}, \bar{Y}_3 - \hat{\beta}_{31} \bar{Y}_2 \right) \\ (\hat{\sigma}_u^2, \hat{\sigma}_v^2) &= (m_{Y_1}^2 - \hat{\sigma}_x^2, m_{Y_3}^2 - \hat{\beta}_{31} m_{Y_2 Y_3}). \end{aligned}$$

Again, we see here that $\underline{\omega}$ has an inverse, so applying Theorem 1, we have that $\underline{\beta}$ is identified.

Now let us assume the addition of a second IV, i.e. a fourth variable related to x , but conditionally independent of the other manifest variables. Suppose this IV is an independent, unbiased measure of x , estimated from some other method. So Y_4 is estimated from an independent method than that of Y_2 , and is different and conditionally independent from the initial IV, Y_3 . Therefore, still having the same models and assumptions for Y_1 , Y_2 , Y_3 , and x , we now also have $Y_4 = x + t$, where $t \sim N(0, \sigma_t^2)$. Under these conditions, the vector of model parameters, which we

assume to have a unique solution, is now

$$\underline{\beta} = (\beta_{10}, \beta_{11}, \sigma_e^2, \sigma_u^2, \beta_{30}, \beta_{31}, \sigma_v^2, \sigma_t^2, \mu_x, \sigma_x^2)',$$

where $\dim(\underline{\beta}) = 10$. Further, the joint marginal distribution is now

$$f(y_1, y_2, y_3, y_4) \sim MVN \left(\begin{bmatrix} \mu_{Y_1}, \mu_{Y_2}, \mu_{Y_3}, \mu_{Y_4} \end{bmatrix}', \begin{bmatrix} \sigma_{Y_1}^2 & \sigma_{Y_1 Y_2} & \sigma_{Y_1 Y_3} & \sigma_{Y_1 Y_4} \\ \sigma_{Y_1 Y_2} & \sigma_{Y_2}^2 & \sigma_{Y_2 Y_3} & \sigma_{Y_2 Y_4} \\ \sigma_{Y_1 Y_3} & \sigma_{Y_2 Y_3} & \sigma_{Y_3}^2 & \sigma_{Y_3 Y_4} \\ \sigma_{Y_1 Y_4} & \sigma_{Y_2 Y_4} & \sigma_{Y_3 Y_4} & \sigma_{Y_4}^2 \end{bmatrix} \right),$$

with $\dim(\underline{\omega}) = 14$. We see here that there are again hopes that $\underline{\beta}$ is identified due to the necessary condition that $\dim(\underline{\beta}) \leq \dim(\underline{\omega})$. In fact, since $\dim(\underline{\omega})$ is strictly greater than $\dim(\underline{\beta})$, we must consider Theorem 2. There do exist invertible, proper subvectors of $\underline{\omega}$ such that $\widehat{\underline{\omega}}_1^{-1}(\widehat{\underline{\omega}}(\widehat{\underline{\beta}})) = \widehat{\underline{\beta}}$. One such just identified subvector is given by

$$\begin{aligned} \widehat{\beta}_{11} &= \frac{m_{Y_1 Y_3}}{m_{Y_2 Y_3}} \\ (\widehat{\sigma}_x^2, \widehat{\sigma}_e^2) &= (m_{Y_2 Y_4}, m_{Y_1}^2 - \widehat{\beta}_{11}^2 \widehat{\sigma}_x^2) \\ (\widehat{\mu}_x, \widehat{\beta}_{10}) &= (\overline{Y}_2, \overline{Y}_1 - \widehat{\beta}_{11} \widehat{\mu}_x) \\ (\widehat{\beta}_{31}, \widehat{\beta}_{30}) &= \left(\frac{m_{Y_2 Y_3}}{\widehat{\sigma}_x^2}, \overline{Y}_3 - \widehat{\beta}_{31} \widehat{\mu}_x \right) \\ (\sigma_u^*, \sigma_v^*, \sigma_t^*) &= (m_{Y_2}^* - \sigma_x^*, m_{Y_3}^* - \beta_{31}^* \sigma_x^*, m_{Y_4}^* - \sigma_x^*). \end{aligned}$$

The remaining elements of $\underline{\omega}(\underline{\beta})$, which were not used to arrive at the above estimators, are given by the vector

$$\underline{\omega}_2 = (\mu_{Y_4}, \sigma_{Y_1 Y_2}, \sigma_{Y_1 Y_4}, \sigma_{Y_3 Y_4})'.$$

So we see that there exists an invertible proper subvector, $\underline{\omega}_1$, with $\dim(\underline{\omega}_1) = \dim(\underline{\beta})$, and thus by Theorem 2, we may conclude that $\underline{\beta}$ is identified.

CHAPTER 4

ESTIMATION OF THE G.S.M.E. MODEL

Assuming the parameters of the GSME model are identified, we proceed to develop a method of estimation. Initially in this chapter, we define the three different types of manifest variables potentially included in the GSME model. After discussing the different types of variables possible, we utilize the conditional independence assumption to write the joint distribution function of the manifest variables in terms of the conditional distributions and the distribution of x . Ultimately, we show how the joint distribution can be approximated by an extension of the Latent Class model and, using a method of generalized least squares, how to obtain estimated generalized nonlinear least squares estimates of the GSME model parameters. We do this, for notational convenience, assuming all variables are univariate and satisfy the conditional independence assumption. Then we show an important example; a generalization to handle a multivariate regression for the outcome of interest which consists of at least one group of conditionally dependent variables. The procedure and its concepts extend to the general multivariate case.

We saw in Chapter 3 that a key component in defining a GSME model is the relation between the conditional distribution parameters of the manifest variables (i.e., the $\underline{\theta}_r$ parameters) and the model parameters (i.e., the $\underline{\beta}_r$ parameters) describing the relation between the parameters modeled (i.e., the $\underline{\gamma}_r$ parameters) for each manifest variable and x . These relations are defined by the mappings in Equation 3.4. Recall that the relationships are defined by the inverse mappings $\underline{\theta}_r = T_r^{-1}(\underline{h}_r(x; \underline{\beta}_r))$. We shall discuss the procedure used to estimate the model parameters, $\underline{\beta} = (\underline{\beta}'_1, \underline{\beta}'_2, \dots, \underline{\beta}'_p, \underline{\theta}'_x)'$. The broad applicability of the GSME model is due, in part, to the fact that it can handle situations involving any type

of variable. This means the model may be applied to all types of categorical and non-categorical data. We distinguish between continuous, categorical variables following a multinomial distribution (which are discrete variables taking on a finite number of values), and “other” discrete variables, i.e., those which do not follow a multinomial distribution but take on discrete values.

In Chapter 3, each $\underline{Y}_r$ had dimension p_r . One such vector could possibly contain all three types of variables: continuous, discrete quantitative, and multinomial. Let $\underline{Y}_{(c)r}$ denote the vector of continuous variables contained in the r th manifest vector of variable, $\underline{Y}_r$, $\underline{Y}_{(d)r}$ the vector of discrete quantitative variables contained in $\underline{Y}_r$, and $\underline{Y}_{(m)r}$ the vector of multinomial-categorical variables contained in $\underline{Y}_r$. This is to allow $\underline{Y}_r$ to range over all possibilities from being a univariate variable of one particular type to a vector of all three types of variables.

To describe the joint marginal distribution of the $\underline{Y}_r$, $r = 1, 2, \dots, p$, as that given at the beginning of Section 3.3 of Chapter 3, we initially define $f_r(\underline{y}_r | x, \underline{\theta}_r)$ as a generic *conditional distribution* for $\underline{Y}_r$. So, for the conditional distribution of $\underline{Y}_r$, we have

$$\begin{aligned} f_r(\underline{y}_r | x, \underline{\theta}_r) &= f_r \left(\underline{y}_{(c)r} | \underline{y}_{(d)r}, \underline{y}_{(m)r}, x, \underline{\theta}_{(c)r} \right)^{I(\text{continuous})} \\ &\quad \times p_r \left(\underline{y}_{(d)r} | \underline{y}_{(m)r}, x, \underline{\theta}_{(d)r} \right)^{I(\text{discrete})} \\ &\quad \times P \left(\underline{Y}_{(m)r} = \underline{y}_{(m)r} | x, \underline{\theta}_{(m)r} \right)^{I(\text{multinomial})}, \end{aligned}$$

where $f_r \left(\underline{y}_{(c)r} | \underline{y}_{(d)r}, \underline{y}_{(m)r}, x, \underline{\theta}_{(c)r} \right)$ is a multivariate continuous density for $\underline{Y}_{(c)r}$, $p_r \left(\underline{y}_{(d)r} | \underline{y}_{(m)r}, x, \underline{\theta}_{(d)r} \right)$ is a multivariate discrete mass function for $\underline{Y}_{(d)r}$, $P \left(\underline{Y}_{(m)r} = \underline{y}_{(m)r} | x, \underline{\theta}_{(m)r} \right)$ are multinomial probabilities defined by the cross-classification of the elements of $\underline{Y}_{(m)r}$, having probabilities defined by the cell probabilities of the cross-classification of its elements, and $I(\cdot)$ are indicators defined to be one, if $\underline{Y}_r$ consists of the type of variable in parentheses, or zero otherwise.

Under the assumption of conditional independence of the $\underline{Y}_r$ given x , $r = 1, 2, \dots, p$, we can write the joint marginal distribution of the manifest variables as:

$$f(\underline{y}_1, \underline{y}_2, \dots, \underline{y}_p) = \int_x \prod_{r=1}^p f_r(\underline{y}_r | x, \underline{\theta}_r) f(x | \underline{\theta}_x) dx, \quad (4.1)$$

where $\int_x$ denotes integration over x . The conditional distribution of the outcome variable of interest, $\underline{Y}_{r^*}$, may be any of the conditional distributions on the right hand side of Equation 4.1, therefore we know that $r^* \in \{1, 2, \dots, p\}$. In attempting to use Equation 4.1 for parameter estimation, a problem arises because the joint distribution on the left hand side is almost always unknown and the integral on the right hand side often cannot be evaluated. Therefore, we must develop a methodology based on Equation 4.1, in order to derive estimates for the model parameters. The notation to invoke this estimation procedure becomes very cumbersome when the model is written in its full generality, as in Equation 4.1. Therefore, for ease of discussion, we show in detail the estimation procedure assuming all variables are univariate and that there are no violations of the conditional independence assumption. Thus, there is no need for the grouping of conditionally dependent variables.

After showing the entire estimation process for the all univariate case, we show a generalizing example of the GSME model where the outcome of interest requires multivariate regression for a group of continuous variables.

The estimation procedure, in summary, amounts to categorizing all variables which are not initially categorical and developing the joint distribution for the resulting categorical variables. In doing so, it is seen that the analog to Equation 4.1 becomes a form of the Latent Class model and the joint distribution after categorization follows a multinomial distribution.

4.1 The Univariate Case

Assume, for notational convenience, that all manifest variables are univariate and satisfy the model assumption of conditional independence given x . Without loss of generality, arrange the p manifest variables such that Y_r , $1 \leq r \leq p'$, are the continuous variables, Y_r , $p' + 1 \leq r \leq p''$, are the “other” discrete variables, and Y_r , $p'' + 1 \leq r \leq p$, are the categorical variables that follow multinomial distributions.

It is useful, when describing the estimation procedure, to define more distinctive notation for the conditional distributions of the three types of manifest variables, instead of letting all of them be represented generically by $f_r(y_r | x, \underline{\theta}_r)$, $r = 1, 2, \dots, p$. Suppose that the form of the distribution function for x is known and represented by $f(x | \underline{\theta}_x)$. Suppose further that the continuous manifest variables have conditional distribution functions with known form denoted by $f_r(y_r | x, \underline{\theta}_r)$, for $1 \leq r \leq p'$, where $\underline{\theta}_r \in \Theta_r$ are $k_r \times 1$ vectors of parameters for each conditional distribution. Suppose that the discrete quantitative variables have known probability mass functions denoted by $p_r(y_r | x, \underline{\theta}_r)$, for $p' + 1 \leq r \leq p''$, where, again, $\underline{\theta}_r \in \Theta_r$ are $k_r \times 1$ vectors of conditional distribution parameters. The remaining variables, Y_r , $p'' + 1 \leq r \leq p$, are “categorical” each following a multinomial distribution. For these variables, we let Y_r have C_r categories, with known conditional cell probabilities given by $P(Y_r = j_r | x, \underline{\theta}_r)$, $j_r = 1, 2, \dots, C_r$. The dimension of the column vector, $\underline{\theta}_r$, of conditional cell probabilities is $k_r = C_r - 1$ and it lies in a parameter space Θ_r .

4.1.1 Translation to Latent Class Model

The first step towards estimation of the parameters in the GSME model is to categorize all continuous and discrete variables and then derive an equation analogous to Equation 4.1, the joint marginal distribution, for categorical variables. Here we describe the categorization method and discuss important notation. First, categorize each continuous Y_r into C_r mutually exclusive and exhaustive categories.

For each Y_r define *intervals* (l_{rj_r}, u_{rj_r}) , $r = 1, 2, \dots, p'$ and $j_r = 1, 2, \dots, C_r$.

Denote the interval defining the j_r th category as I_{rj_r} and the $\text{length}(I_{rj_r})$ by L_{rj_r} .

To indicate the probability that a continuous manifest variable falls into a certain interval, we write $P(Y_r \in I_{rj_r})$. Note the intervals or "categories" need not be of equal length.

The "other" discrete variables, i.e., the non-continuous variables not following a multinomial distribution, may also need to be categorized. The process by which the Y_r , $p' + 1 \leq r \leq p''$, are categorized, is essentially the same as that for the continuous manifest variables. For these discrete variables requiring categorization I_{rj_r} may be thought of as a collection of contiguous values of the discrete variable. It is possible in an applied context, that a discrete variable of this type may take on such a small number of values, that no further categorization is needed. In these cases each I_{rj_r} would actually represent a single value, as opposed to the collection of discrete values. We shall denote the probability that a discrete manifest variable falls into the set of values I_{rj_r} as $P(Y_r \in I_{rj_r})$.

The discrete manifest random variables, that follow a multinomial distribution are categorical variables that require no further categorization. Therefore Y_r , $p'' + 1 \leq r \leq p$, remain unchanged. (If collapsing of categories is required it should be done *a priori* and, hence, is already reflected in definitions of Y_r , $r = p'' + 1, \dots, p$.) Recall that the total number of categories of Y_r is denoted by C_r , $p'' + 1 \leq r \leq p$, and that these variables have conditional cell probabilities given by $P(Y_r = j_r \mid x, \underline{\theta}_r)$, where $j_r = 1, 2, \dots, C_r$. Since nothing changes for the categorical manifest variables after the categorization step, the notation for the conditional cell probabilities remains the same as used previously.

Let x be categorized in a similar manner as mentioned above for the continuous manifest variables, where its categories are indexed by the subscript j_x . The categories are denoted by I_{j_x} , where $\text{length}(I_{j_x}) = L_{j_x}$, $j_x = 1, 2, \dots, C_x$, and the

probability that x falls into one of its intervals is denoted $P(x \in I_{j_x})$. Let x_{j_x} represent the midpoint of interval j_x .

An important note, now that the categorization has been discussed, is that the difference between the “discrete” and “categorical” manifest variables as discussed at the beginning of this chapter is actually intended to mean the difference between non-continuous variables requiring further categorization and non-continuous variables where no modification is needed. By this we mean that there may be scenarios, depending on sample size perhaps, in which a “discrete” manifest variable may actually require no categorization or a “categorical” manifest variable may require further categorization. As previously alluded to, if a discrete variable needed no further categorization, then the I_{rj_x} for that variable would represent a single value of that variable. This may be needed in instances where a discrete variable takes on a small number of values. Further categorizing a categorical variable would entail combining possible realizations of that variable and defining new categories, and thus new conditional cell probabilities, based on groups of the original categories, as mentioned above. This may be required if there are an unusually large number of categories for a categorical variable. We use the “discrete” and “categorical” terminology because in practice, discrete variables will generally be those that require categorization, and categorical variables will generally be those that will not require further modification.

The joint distribution on the left hand side of Equation 4.1 after categorization is a multinomial distribution. For the categorized variables, Equation 4.1 becomes

$$\begin{aligned}
\pi_{j_1 j_2 \dots j_p} &= P(Y_1 \in I_{j_1}, \dots, Y_{p''} \in I_{p'' j_{p''}}, Y_{p''+1} = j_{p''+1}, \dots, Y_p = j_p) \\
&= \sum_{j_x} \prod_{r=1}^{p''} P(Y_r \in I_{r j_r} \mid x \in I_{j_x}, \underline{\beta}_r) \\
&\quad \times \prod_{r=p''+1}^p P(Y_r = j_r \mid x \in I_{j_x}, \underline{\beta}_r) P(x \in I_{j_x} \mid \underline{\theta}_x), \\
&= \pi_{j_1 j_2 \dots j_p}(\underline{\beta}),
\end{aligned} \tag{4.2}$$

where $\pi_{j_1 j_2 \dots j_p}$ indicates the probability of falling into multinomial cell $j_1 j_2 \dots j_p$, for $j_r = 1, 2, \dots, C_r$, $r = 1, 2, \dots, p$. The last equality is to note that we write these cell probabilities as functions of the model parameters, $\underline{\beta}$, again, by substitution of the distributional parameters for the mappings defined in Equation 3.4 which are in terms of x and $\underline{\beta}_r$, i.e., $\underline{\theta}_r = \underline{T}_r^{-1}(\underline{h}_r(x; \underline{\beta}_r))$, for all r , when calculating the cell probabilities. This equation is a variation of the classical Latent Class model which was discussed in detail by Lazarsfeld and Henry [38]. *Latent Class* models are those which may be applied to cross tabulated data. The idea is that a possibly complex relationship between observed variables can be explained by a simple relationship that holds for unobservable latent variables. The data for Latent Class models are characteristics, discrete-valued variables, that are assumed to be indicators of some underlying, unobservable concept or latent variable [63, p. 194]. In other words, a Latent Class model can be thought of as an expression which includes at least one covariate that may be considered to be measured with error. In the simplest form of the Latent Class model, the observable variables are dichotomous.

Note that in Equation 4.2 the integral over x has become a summation over j_x and so Equation 4.2 can be seen as a Latent Class model. The model parameters enter nonlinearly in the right side of Equation 4.2 through the conditional probabilities which are calculated for x by $P(x \in I_{j_x} \mid \underline{\theta}_x) = \int_{I_{j_x}} f(x \mid \underline{\theta}_x) dx \doteq f(x_{j_x} \mid$

$\underline{\theta}_x)L_{j_x}$, for I_{j_x} of small length, where, recall that $\text{length}(I_{j_x}) = L_{j_x}$. (Note that if any particular type of the three kinds of manifest variables does not exist in an applied context, assume a value of unity throughout the following methodology for that variables' density and ignore the associated integration or summation.) For the cell probabilities of a continuous Y_r , we have

$$\begin{aligned}
 P(Y_r \in I_{rj_r} \mid x \in I_{j_x}, \underline{\beta}_r) &= \frac{P(Y_r \in I_{rj_r}, x \in I_{j_x} \mid \underline{\beta}_r)}{P(x \in I_{j_x} \mid \underline{\theta}_x)} \\
 &= \frac{\int_{I_{rj_r}} \int_{I_{j_x}} f_r(y_r, x \mid \underline{\theta}_{rx}) dx dy_r}{\int_{I_{j_x}} f(x \mid \underline{\theta}_x) dx} \\
 &\doteq \frac{\int_{I_{rj_r}} f_r(y_r, x_{j_x} \mid \underline{\theta}_{rx}) L_{j_x} dy_r}{f(x_{j_x} \mid \underline{\theta}_x) L_{j_x}} \\
 &= \int_{I_{rj_r}} f_r(y_r \mid x = x_{j_x}, \underline{\beta}_r) dy_r,
 \end{aligned}$$

for I_{j_x} of small length, dy_r corresponding to the integration over I_{rj_r} , and, where $\underline{\theta}_{rx}$ are the (unknown) parameters of the joint distribution of Y_r and x . If Y_r is discrete the probabilities are

$$P(Y_r \in I_{rj_r} \mid x_{j_x} \in I_{j_x}, \underline{\beta}_r) = \sum_{y_r \in I_{rj_r}} p_r(y_r \mid x = x_{j_x}, \underline{\beta}_r),$$

where the summation is over discrete values defining the set of values, I_{rj_r} .

In practice, it is during these previous calculations that the model parameters, i.e., the elements of $\underline{\beta} = (\underline{\beta}'_1, \underline{\beta}'_2, \dots, \underline{\beta}'_p, \underline{\theta}'_x)'$, enter the right hand side of Equation 4.2 by substituting the conditional distributional parameters of the conditional densities with the functions of these model parameters whose mappings were defined in Equation 3.4. We write $f_r(y_r \mid x = x_{j_x}, \underline{\beta}_r)$, but it should be understood that this notation represents $f_r(y_r \mid x = x_{j_x}, \underline{T}_r^{-1} \underline{h}_r(x_{j_x}; \underline{\beta}_r))$, which of course equals $f_r(y_r \mid x = x_{j_x}, \underline{\theta}_r)$. The notation for the non-continuous variables, $p_r(y_r \mid x = x_{j_x}, \underline{\beta}_r)$, should be interpreted similarly.

In its full generality, for univariate manifest variables, Equation 4.2 becomes

$$\begin{aligned}
 \pi_{j_1 j_2 \dots j_p} &\doteq \sum_{j_z} \prod_{r=1}^{p'} \left[\int_{I_{rj_r}} f_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) dy_r \right] \\
 &\quad \prod_{r=p''+1}^{p''} \left[\sum_{y_r \in I_{rj_r}} p_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) \right] \\
 &\quad \prod_{r=p'''+1}^p P(Y_r = j_r \mid x = x_{j_z}, \underline{\beta}_r) f(x_{j_z} \mid \underline{\theta}_x) L_{j_z} \quad (4.3) \\
 &= \pi_{j_1 j_2 \dots j_p}(\underline{\beta}),
 \end{aligned}$$

where $\pi_{j_1 j_2 \dots j_p}(\underline{\beta})$ indicates the cell probability written in terms of the model parameters. Essentially, then, if the r th variable is continuous, then $\int_{I_{rj_r}} f_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) dy_r$ is included in the product and if it is discrete requiring categorization, then $\sum_{y_r \in I_{rj_r}} p_r(y_r \mid x = x_{j_z}, \underline{\beta}_r)$ is included. Again, note that if there are no variables of a given type (continuous, discrete, or categorical), then the factor associated with that type of variable is dropped from Equation 4.3.

4.1.2 Estimation of Model Parameters

In order to estimate the parameters of the extended Latent Class model defined by Equation 4.3, we note that the sample proportions of the data falling into the categories indicated on the left side of Equation 4.3, i.e., the $(N - 1) \times 1$ vector $\hat{\underline{\pi}}$, are the maximum likelihood estimators of $P(Y_1 \in I_{1j_1}, \dots, Y_{p''} \in I_{p''j_{p''}}, Y_{p''+1} = j_{p''+1}, \dots, Y_p = j_p)$, with corresponding $(N - 1)$ -dimensional vector of cell probabilities $\underline{\pi}$, $N = \prod_{r=1}^p C_r$. In the estimation procedure, the last element of $\underline{\pi}$ and $\hat{\underline{\pi}}$ must be dropped to avoid a singularity in the variance-covariance matrix, hence the dimensions of those column vectors are $N - 1$. Due to the invariance property of maximum likelihood estimators, substituting $\hat{\underline{\pi}}$ on the left hand side of Equation 4.3 and solving for the model parameters, $\underline{\beta}$, on the right hand side as functions of the elements of $\hat{\underline{\pi}}$ gives solutions that are maximum likelihood estimates. This is possible when $N - 1 = \dim(\underline{\beta})$. If $N - 1 > \dim(\underline{\beta})$, then there is

no exact solution to the equations $\hat{\pi}_{j_1 j_2 \dots j_p}$ equal to the right hand side of Equation 4.3, and iterative generalized nonlinear least squares, similar to that proposed by Shah [52] for diagnostic tests in the usual Latent Class models, can be used to obtain least squares solutions.

Here, we will give a brief motivation for the use of the method of estimation developed by Shah [52], which was based on Estimated Generalized Nonlinear Least Squares (henceforth, EGNLS) in the special case where Y_r were binary variables for diagnostic tests and no covariates were involved. We generalize this method for use in solving GSME models. EGNLS amounts to maximization of an approximating multivariate normal likelihood. The resulting estimates are shown, in subsequent sections, to be weakly consistent and asymptotically normal. The estimation procedure is developed from the initial relationship presented in Equation 4.3 and concisely summarized in vector form as

$$\hat{\underline{\pi}} = \underline{\pi}(\underline{\beta}) + \underline{\varepsilon}, \quad (4.4)$$

where $\underline{\beta} = (\underline{\beta}'_1, \underline{\beta}'_2, \dots, \underline{\beta}'_p, \underline{\theta}'_x)'$ is the $(l_1 + l_2 + \dots + l_p + k_x)$ -dimensional column vector of model parameters and $\hat{\underline{\pi}}$ is the vector of sample proportions which are unbiased and consistent maximum likelihood estimates of the vector of the multinomial cell probabilities, $P(Y_1 \in I_{1j_1}, \dots, Y_{p''} \in I_{p''j_{p''}}, Y_{p''+1} = j_{p''+1}, \dots, Y_p = j_p) = \pi_{j_1 j_2 \dots j_p}$, for all possible combinations of the categorized manifest variables. Further, the multinomial cell probabilities, $\underline{\pi}(\underline{\beta})$, may be written as nonlinear functions of the model parameters, as defined by the right hand side of Equation 4.3. Recall that

from Equation 4.3, we have

$$\begin{aligned} \pi_{j_1 j_2 \dots j_p}(\underline{\beta}) &= \sum_{j_z} \prod_{r=1}^{p'} \left[\int_{I_{r j_r}} f_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) dy_r \right] \\ &\quad \prod_{r=p'+1}^{p''} \left[\sum_{y_r \in I_{r j_r}} p_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) \right] \\ &\quad \prod_{r=p''+1}^p P(Y_r = j_r \mid x = x_{j_z}, \underline{\beta}_{\underline{r}}) f(x_{j_z} \mid \underline{\theta}_{\underline{z}}) L_{j_z}, \end{aligned}$$

where $\pi_{j_1 j_2 \dots j_p}(\underline{\beta})$ is the cell probability associated with the category in the $(j_1, j_2, \dots, j_p)$ th cell of the cross-classification of the categorized manifest variables, where $j_r = 1, 2, \dots, C_r$, for all r . In Equation 4.4, $\widehat{\underline{\pi}}$ is the vector of estimates of the $\pi_{j_1 j_2 \dots j_p}$ and $\underline{\pi}(\underline{\beta})$ is the vector of corresponding cell probabilities which are written as nonlinear functions of the model parameters. We further have that $E(\underline{\varepsilon}) = \underline{0}$ and $Var(\underline{\varepsilon}) = n^{-1}V$, where $V = Cov(\underline{U}_i)$ and $\underline{U}_i$ denotes an observation vector on the i th subject with the l th element defined by $U_{il} = 1$ if the subject falls into cell l of $\underline{\pi}$ from the joint multinomial distribution, and $U_{il} = 0$ otherwise, $l = 1, 2, \dots, N - 1$, $N = \prod_{r=1}^p C_r$. Therefore, $V = diag(\underline{\pi}(\underline{\beta})) - \underline{\pi}(\underline{\beta})\underline{\pi}(\underline{\beta})'$. Because of the linear dependency $\sum_l \widehat{\pi}_l = 1$, V would be a singular matrix if the last element of $\underline{\pi}$ was not dropped. In the estimation procedure, recall, the N th component of $\widehat{\underline{\pi}}$, $\underline{U}_i$, and $\underline{\pi}$ is deleted so that V is an $(N - 1) \times (N - 1)$ positive definite matrix.

Generalized nonlinear least squares estimates of $\underline{\beta}$ are obtained by minimizing the quadratic form

$$Q(\underline{\beta}, V) = n (\widehat{\underline{\pi}} - \underline{\pi}(\underline{\beta}))' V^{-1} (\widehat{\underline{\pi}} - \underline{\pi}(\underline{\beta})),$$

with respect to $\underline{\beta}$. Estimation is hindered by the fact that V is unknown and depends on the parameters $\underline{\beta}$. EGNLS amounts to minimizing

$$Q(\underline{\beta}, \widehat{V}) = n (\widehat{\underline{\pi}} - \underline{\pi}(\underline{\beta}))' \widehat{V}^{-1} (\widehat{\underline{\pi}} - \underline{\pi}(\underline{\beta})), \quad (4.5)$$

where $\hat{V}$ is any calculated, consistent, positive definite estimator of V , and therefore is not a function of the parameters. This requirement for $\hat{V}^{-1}$ should hold as long as there are no zero counts in any cell.

The solutions which minimize this quadratic are obtained by setting its derivative, with respect to $\underline{\beta}'$, equal to a vector of zeros and solving for the elements of $\underline{\beta}$. For the GSME model this is no trivial task, due to the difficulty in deriving the first derivative of Equation 4.5. A further difficulty arises from the need to use the second derivative of $Q(\underline{\beta}, \hat{V})$ in developing the asymptotic properties of the EGNLS estimators of the GSME model. The difficulties arise, in part, due to the numerous and complex derivatives of matrices and vectors, but also in differentiating the compound mappings involved in $\underline{\pi}(\underline{\beta})$. Before we derive the first derivative of Equation 4.5, we state and prove a general chain rule for derivatives of multiply compounded, multivariable functions, which has not previously been written, to this author's knowledge.

Recall that $\underline{\pi}(\underline{\beta})$ involves the vectors $\underline{T}^{-1}$ and $\underline{h}$ which are defined to be column vectors whose elements are the individual mappings used to relate the parameters of the conditional distribution of $\underline{Y}_r$ given x to x . When written in terms of $\underline{\beta}$, we abbreviate the vector of multinomial probabilities as $\underline{\pi}(\underline{\beta})$ where it should be understood that this is a symbolic representation of the entire vector $\underline{\pi}(\underline{T}^{-1}(\underline{h}(x; \underline{\beta})))$. Schott [49, p. 327] established a rule for the derivative of a compound function that is vector valued. Using his notation, he states that if y and g are real valued functions, $\underline{f}$ is a vector of functions, and $\underline{x}$ is a vector such that $y(\underline{x}) = g(\underline{f}(\underline{x}))$, then

$$\begin{aligned} \frac{\partial}{\partial x_i} y(\underline{x}) &= \sum_j \left(\frac{\partial}{\partial f_j} g(\underline{f}) \right) \left(\frac{\partial}{\partial x_i} f_j(\underline{x}) \right) \\ &= \left(\frac{\partial}{\partial \underline{f}'} g(\underline{f}) \right) \left(\frac{\partial}{\partial x_i} \underline{f}(\underline{x}) \right), \end{aligned} \quad (4.6)$$

for $i = 1, 2, \dots, n$. These derivatives with respect to x_i can be placed into one vector, $\frac{\partial y(\underline{x})}{\partial \underline{x}}$, which can be written as

$$\frac{\partial}{\partial \underline{x}'} y(\underline{x}) = \left(\frac{\partial}{\partial \underline{f}'} g(\underline{f}) \right) \left(\frac{\partial}{\partial \underline{x}'} \underline{f}(\underline{x}) \right).$$

We generalize this to multiply compounded functions of F vector valued functions in the following result.

Result: Let $y(\underline{x}) = g(\underline{f}_F \dots (\underline{f}_2(\underline{f}_1(\underline{x}))))$ be a general *multiply compounded, multivariable function*. Then the derivative of $y(\underline{x})$ with respect to $\underline{x}'$ can be written as

$$\frac{\partial}{\partial \underline{x}'} y(\underline{x}) = \left(\frac{\partial}{\partial \underline{f}_F'} g(\underline{f}_F) \right) \left(\frac{\partial}{\partial \underline{f}_{F-1}'} \underline{f}_F(\underline{f}_{F-1}) \right) \dots \left(\frac{\partial}{\partial \underline{f}_2'} \underline{f}_3(\underline{f}_2) \right) \left(\frac{\partial}{\partial \underline{f}_1'} \underline{f}_2(\underline{f}_1) \right) \left(\frac{\partial}{\partial \underline{x}'} \underline{f}_1(\underline{x}) \right).$$

Proof: By Schott's result, the result holds for $F = 1$. That is, we may write

$$\frac{\partial}{\partial x_i} y(\underline{x}) = \sum_j \left(\frac{\partial}{\partial f_{1j}} g(\underline{f}_1) \right) \left(\frac{\partial}{\partial x_i} f_{1j}(\underline{x}) \right)$$

and

$$\frac{\partial}{\partial \underline{x}'} y(\underline{x}) = \left(\frac{\partial}{\partial \underline{f}_1'} g(\underline{f}_1) \right) \left(\frac{\partial}{\partial \underline{x}'} \underline{f}_1(\underline{x}) \right).$$

Further, for the j th element of $\underline{f}_1$, f_{1j} , we have

$$\frac{\partial}{\partial f_{1j}} g(\underline{f}_2(\underline{f}_1)) = \sum_k \left(\frac{\partial}{\partial f_{2k}} g(\underline{f}_2) \right) \left(\frac{\partial}{\partial f_{1j}} f_{2k}(\underline{f}_1) \right),$$

by Equation 4.6. Combining these gives,

$$\begin{aligned} \frac{\partial}{\partial x_i} g(\underline{f}_2(\underline{f}_1(\underline{x}))) &= \sum_j \sum_k \left(\frac{\partial}{\partial f_{2k}} g(\underline{f}_2) \right) \left(\frac{\partial}{\partial f_{1j}} f_{2k}(\underline{f}_1) \right) \left(\frac{\partial}{\partial x_i} f_{1j}(\underline{x}) \right) \\ &= \left(\frac{\partial}{\partial \underline{f}_2'} g(\underline{f}_2) \right) \left(\frac{\partial}{\partial \underline{f}_1'} \underline{f}_2(\underline{f}_1) \right) \left(\frac{\partial}{\partial x_i} \underline{f}_1(\underline{x}) \right) \end{aligned}$$

as the i th element of the row vector $\frac{\partial}{\partial \underline{x}} g(\underline{f}_2(\underline{f}_1(\underline{x})))$. This process may be repeated F times giving the result that if $y(\underline{x}) = g(\underline{f}_F \dots (\underline{f}_2(\underline{f}_1(\underline{x}))))$, then

$$\frac{\partial}{\partial \underline{x}'} y(\underline{x}) = \left(\frac{\partial}{\partial \underline{f}'_F} g(\underline{f}_F) \right) \dots \left(\frac{\partial}{\partial \underline{f}'_2} \underline{f}_3(\underline{f}_2) \right) \left(\frac{\partial}{\partial \underline{f}'_1} \underline{f}_2(\underline{f}_1) \right) \left(\frac{\partial}{\partial \underline{x}'} \underline{f}_1(\underline{x}) \right),$$

which is arrived at by repeatedly replacing the summations with matrix multiplication, as was done above. $\diamond$

To apply this general result for multiply compounded, multivariable functions to the GSME model we let $y(\underline{x}) = \pi_j(\underline{\beta})$ and $g(\underline{f}_2(\underline{f}_1(\underline{x}))) = \pi_j(\underline{T}^{-1}(\underline{h}(\underline{x}; \underline{\beta})))$, for $j = 1, 2, \dots, N$, $N = \prod_{r=1}^p C_r$. Placing the resulting row vectors of the derivative of $\pi_j(\underline{T}^{-1}(\underline{h}(\underline{x}; \underline{\beta})))$ with respect to $\underline{\beta}'$, one atop another, gives us the $N - 1 \times \nu$ matrix

$$\frac{\partial}{\partial \underline{\beta}'} \underline{\pi}(\underline{\beta}) = \frac{\partial}{\partial \underline{\beta}'} \underline{\pi}(\underline{T}^{-1}(\underline{h}(\underline{x}; \underline{\beta}))),$$

where ν is the dimension of $\underline{\beta}$ and $N - 1$ is the dimension of $\underline{\pi}(\underline{\beta})$, due to the need to drop the N th element to avoid the linear dependency which is created because the cell probabilities sum to one. In general, using the result proven above, the (j, i) th element of $\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'}$ is

$$\left(\frac{\partial}{\partial \underline{T}^{-1'}} \pi_j(\underline{T}^{-1}) \right) \left(\frac{\partial}{\partial \underline{h}'} \underline{T}^{-1}(\underline{h}) \right) \left(\frac{\partial}{\partial \underline{\beta}_i} \underline{h}(\underline{x}; \underline{\beta}) \right),$$

where the last k_x of the $\underline{\beta}_i$ should be understood to simply be symbolic representations of the elements of $\underline{\theta}_x$. This derivative will be used in what follows, especially later in the chapter in developing the asymptotic properties of the estimators.

We will simply be writing $\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'}$ to represent the entire $N - 1 \times \nu$ matrix of the derivatives of the multiply compounded, multivariable vector valued function, $\underline{\pi}(\underline{\beta})$.

Let us recall that we are trying to minimize the quadratic form $Q(\underline{\beta}, \hat{V})$ given in Equation 4.5. To write out the derivative required to perform this minimization, we will need to apply the result proven above along with another result from Schott [49] concerning the derivative of a quadratic form of a vector $\underline{x}$ and a symmetric

matrix A . Schott [49, p. 329-330] shows that $\frac{\partial}{\partial \underline{x}'} \underline{x}' A \underline{x} = 2 \underline{x}' A$. Applied to our quadratic form and letting $(\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta})) = \Delta \underline{\pi}$, we have

$$\frac{\partial}{\partial \Delta \underline{\pi}'} Q(\underline{\beta}, \hat{V}) = 2n (\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta}))' \hat{V}^{-1}$$

and therefore

$$\frac{\partial}{\partial \underline{\beta}'} Q(\underline{\beta}, \hat{V}) = \left(\frac{\partial}{\partial \Delta \underline{\pi}'} Q(\underline{\beta}, \hat{V}) \right) \left(\frac{\partial}{\partial \underline{\pi}(\underline{\beta})'} \Delta \underline{\pi} \right) \left(\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'} \right).$$

Because $\left(\frac{\partial}{\partial \underline{\pi}(\underline{\beta})'} \Delta \underline{\pi} \right) = -I$, where I is the identity matrix, we have

$$\frac{\partial}{\partial \underline{\beta}'} Q(\underline{\beta}, \hat{V}) = -2n (\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta}))' \hat{V}^{-1} \left(\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'} \right).$$

Again, it is known that setting this derivative equal to a vector of zeros and solving for $\underline{\beta}$ minimizes the quadratic form in Equation 4.5, i.e., solving $\frac{\partial}{\partial \underline{\beta}'} Q(\underline{\beta}, \hat{V}) = \underline{0}'$ for $\underline{\beta}$, provided that only a global minimum exists. This solution, however, is the same as the solution when using the estimating equations $\underline{S}_n(\underline{\beta}) = \left(\frac{\partial}{\partial \underline{\beta}'} Q(\underline{\beta}, \hat{V}) \right)' / (-2n) = \underline{0}$. Therefore, we solve the following system of equations:

$$\underline{S}_n(\underline{\beta}) \equiv \left(\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'} \right)' \hat{V}^{-1} (\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta})) = \underline{0}. \quad (4.7)$$

In the special case where the Y_r , $r = 1, 2, \dots, p$, are indicators (0,1) of the results (no disease, disease) of p diagnostic tests, Shah [52] suggests using sample proportions, $\hat{\underline{\pi}}$, to estimate V consistently. The Continuous Mapping Theorem [51, p. 24] states that a continuous function of a consistent estimate is consistent for the same function of the consistent estimate's probability limit, if the function is continuous with a probability of one. We know that $V = \text{diag}(\underline{\pi}(\underline{\beta})) - \underline{\pi}(\underline{\beta})\underline{\pi}(\underline{\beta})'$ and $\hat{\underline{\pi}}$ are sample proportions which are unbiased and consistent estimates of $\underline{\pi}(\underline{\beta})$, therefore, our consistent estimator of V shall be $\hat{V} = \text{diag}(\hat{\underline{\pi}}) - \hat{\underline{\pi}}\hat{\underline{\pi}}'$, provided all the elements of $\hat{\underline{\pi}}$ are strictly between zero and one. This is to guarantee

that $\hat{V}$ is positive definite. Another possible choice as an estimate of V could be $\text{diag}(\pi(\hat{\beta})) - \pi(\hat{\beta})\pi(\hat{\beta})'$, where $\hat{\beta}$ is any consistent estimate of β and, again, provided that the elements of $\pi(\hat{\beta})$ are greater than zero, but less than one, to ensure that $\hat{V}$ is positive definite.

A widely used method for computing nonlinear least squares is the modified Gauss-Newton method, which will be used to estimate the β parameters on the right hand side of Equation 4.4 (see, for example, [31]). The reason for using the *modified* Gauss-Newton procedure, is because it is known that a drawback to using the standard Gauss-Newton procedure, is that there is no guarantee that the standard algorithm will converge. In fact, Wolfe [64, p. 230] states that “A major difficulty with the Gauss-Newton Method is the initial estimate is seldom sufficiently close to ensure convergence to the true value.” Further, the modified Gauss-Newton method, “whilst sharing the advantageous features of the Gauss-Newton method, has the additional merits of a guaranteed convergence” [31, p. 269]. Hence, we utilize the modified Gauss-Newton procedure.

The first step in deriving the EGNLS estimates is to substitute a first order Taylor series approximation about consistent, initial/trial values, $\hat{\beta}_T$, to the nonlinear function $\pi(\beta)$ in the quadratic form $Q(\beta, \hat{V})$. The approximation is $\pi(\beta) \approx \pi(\hat{\beta}_T) + \left(\frac{\partial \pi(\hat{\beta}_T)}{\partial \beta'} \right) (\beta - \hat{\beta}_T)$, where throughout $\frac{\partial \pi(\hat{\beta}_T)}{\partial \beta'}$ should be interpreted as the derivative of $\pi(\beta)$ with respect to β' evaluated at the initial estimate $\hat{\beta}_T$, i.e., $\frac{\partial \pi(\hat{\beta}_T)}{\partial \beta'} = \frac{\partial \pi(\beta)}{\partial \beta'} \Big|_{\beta=\hat{\beta}_T}$. Using this approximation, it can be shown (see, for example, [20, p. 221]) that the approximating quadratic form is minimized

$$\hat{\beta}_M = \hat{\beta}_T + \left[\left(\frac{\partial \pi(\hat{\beta}_T)}{\partial \beta'} \right)' \hat{V}^{-1} \left(\frac{\partial \pi(\hat{\beta}_T)}{\partial \beta'} \right) \right]^{-1} \left(\frac{\partial \pi(\hat{\beta}_T)}{\partial \beta'} \right)' \hat{V}^{-1} [\hat{\pi} - \pi(\hat{\beta}_T)]. \quad (4.8)$$

There is no guarantee that $\hat{\beta}_M$ is better than $\hat{\beta}_T$ for approximating the least squares estimate, $\hat{\beta}$, in the sense that $Q(\hat{\beta}_M, \hat{V}) < Q(\hat{\beta}_T, \hat{V})$. However, one can

show that, there exists a λ^* such that all points $\widehat{\underline{\beta}}$, defined as $\widehat{\underline{\beta}} = \widehat{\underline{\beta}}_T + \lambda(\widehat{\underline{\beta}}_M - \widehat{\underline{\beta}}_T)$, $0 < \lambda < \lambda^*$, are such that $Q(\widehat{\underline{\beta}}, \widehat{V}) < Q(\widehat{\underline{\beta}}_T, \widehat{V})$. The remaining two steps describe the modified Gauss-Newton method. Letting $\widehat{\underline{\beta}}_k$ denote the estimate after k iterations, compute, at iteration $k + 1$,

$$\underline{D}_k = \left[\left(\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'} \right)' \widehat{V}^{-1} \left(\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'} \right) \right]^{-1} \left(\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'} \right)' \widehat{V}^{-1} [\widehat{\underline{\pi}} - \pi(\widehat{\underline{\beta}}_k)]. \quad (4.9)$$

Next, find a λ_k between 0 and 1 such that $Q(\widehat{\underline{\beta}}_k + \lambda_k \underline{D}_k, \widehat{V}) < Q(\widehat{\underline{\beta}}_k, \widehat{V})$ and set the new estimate as $\widehat{\underline{\beta}}_{k+1} = \widehat{\underline{\beta}}_k + \lambda_k \underline{D}_k$. One can use any reasonable convergence criteria to determine if the algorithm has converged, e.g., based on the distance between estimates at iteration k and $k + 1$.

Here, we develop an expression for $\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'}$ in Equation 4.9 where we do so through repeated use of the chain rule. The result of the derivative is a $(N - 1) \times (l_1 + l_2 + \dots + l_p + k_x)$ matrix of derivatives, i.e., $\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'} = \left[\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'_1}, \frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'_2}, \dots, \frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'_p}, \frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\theta}'_x} \right]$, where $\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'_r} = \left[\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \beta_{r,1}}, \frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \beta_{r,2}}, \dots, \frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \beta_{r,l_r}} \right]$, for all r , and $\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\theta}'_x} = \left[\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \theta_1}, \frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \theta_2}, \dots, \frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \theta_{k_x}} \right]$. The dimension of each $\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\beta}'_r}$ is $(N - 1) \times l_r$ and that of $\frac{\partial \pi(\widehat{\underline{\beta}}_k)}{\partial \underline{\theta}'_x}$ is $(N - 1) \times k_x$.

We shall show these derivatives in terms of a generic, individual β for the r' th variable, i.e., $\beta_{r'i}$, say. This should be interpreted as the i th parameter, $i = 1, 2, \dots, l_{r'}$, for the r' th manifest variable. When the derivative is performed for the parameters of x , we show the derivative with respect to the i th parameter of $\underline{\theta}_x$, i.e., θ_i , where in this case $i = 1, 2, \dots, k_x$. Recall for use in what follows that the abbreviated notation for $\underline{\mathcal{I}}_r^{-1}(\underline{h}_r(x; \underline{\beta}_r))$ is $\underline{\mathcal{I}}_r^{-1}$ and for $\underline{h}_r(x; \underline{\beta}_r)$ the abbreviation is $\underline{h}_r$, for each r .

Given the model in Equation 4.3, the derivative of $\pi_{j_1 j_2 \dots j_p}(\underline{\beta})$, with respect to the i th parameter of interest for the r' th manifest variable is

$$\begin{aligned}
 \frac{\partial \pi_{j_1 j_2 \dots j_p}(\underline{\beta})}{\partial \beta_{r'i}} &= \frac{\partial \sum_{j_z} \text{ABCD}}{\partial \beta_{r'i}} \\
 &= \sum_{j_z} \left[\frac{\partial \text{A}}{\partial \beta_{r'i}} \text{BCD} \right] \text{ if } r' \leq p' \\
 &= \sum_{j_z} \left[\frac{\partial \text{B}}{\partial \beta_{r'i}} \text{ACD} \right] \text{ if } p' + 1 \leq r' \leq p'' \\
 &= \sum_{j_z} \left[\frac{\partial \text{C}}{\partial \beta_{r'i}} \text{ABD} \right] \text{ if } p'' + 1 \leq r' \leq p \\
 &= \sum_{j_z} \left[\frac{\partial \text{D}}{\partial \theta_i} \text{ABC} \right] \text{ if } \underline{\beta}_{r'} = \underline{\theta}_x, \quad (4.10)
 \end{aligned}$$

where

$$\begin{aligned}
 \text{A} &= \prod_{r=1}^{p'} \left[\int_{I_{rj_r}} f_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) dy_r \right] \\
 \text{B} &= \prod_{r=p'+1}^{p''} \left[\sum_{y_r \in I_{rj_r}} p_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) \right] \\
 \text{C} &= \prod_{r=p''+1}^p P(Y_r = j_r \mid x = x_{j_z}, \underline{\beta}_r) \\
 \text{D} &= f(x_{j_z} \mid \underline{\theta}_x) L_{j_z}
 \end{aligned}$$

and for the derivatives, we have

$$\begin{aligned}
 \frac{\partial \text{A}}{\partial \beta_{r'i}} &= \frac{\partial \text{A}}{\partial \int_{I_{r'j_{r'}}} f_{r'}(y_{r'} \mid x = x_{j_z}, \underline{\beta}_{r'}) dy_{r'}} \frac{\partial \int_{I_{r'j_{r'}}} f_{r'}(y_{r'} \mid x = x_{j_z}, \underline{\beta}_{r'}) dy_{r'}}{\partial \underline{T}_{r'}^{-1'}} \\
 &\quad \times \frac{\partial \underline{T}_{r'}^{-1}}{\partial \underline{h}_{r'}} \frac{\partial \underline{h}_{r'}}{\partial \beta_{r'i}} \\
 &= \prod_{r \neq r'} \left[\int_{I_{rj_r}} f_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) dy_r \right] \frac{\partial \int_{I_{r'j_{r'}}} f_{r'}(y_{r'} \mid x = x_{j_z}, \underline{\beta}_{r'}) dy_{r'}}{\partial \underline{T}_{r'}^{-1'}} \\
 &\quad \times \frac{\partial \underline{T}_{r'}^{-1}}{\partial \underline{h}_{r'}} \frac{\partial \underline{h}_{r'}}{\partial \beta_{r'i}}
 \end{aligned}$$

and

$$\begin{aligned}\frac{\partial D}{\partial \theta_i} &= \frac{\partial D}{\partial f(x_{j_x} | \underline{\theta}_x)} \frac{\partial f(x_{j_x} | \underline{\theta}_x)}{\partial \theta_i} \\ &= L_{j_x} \frac{\partial f(x_{j_x} | \underline{\theta}_x)}{\partial \theta_i}.\end{aligned}$$

The calculations for $\frac{\partial B}{\partial \beta_{r'i}}$ and $\frac{\partial C}{\partial \beta_{r'i}}$ are similar as to those for $\frac{\partial A}{\partial \beta_{r'i}}$, accept with $\sum_{y_{r'} \in I_{r'j_{r'}}} p_{r'}(y_{r'} | x = x_{j_x}, \underline{\beta}_{r'})$ and $P(Y_{r'} = j_{r'} | x = x_{j_x}, \underline{\beta}_{r'})$, respectively, replacing $\int_{I_{r'j_{r'}}} f_{r'}(y_{r'} | x = x_{j_x}, \underline{\beta}_{r'}) dy_{r'}$. The final step would be the concatenation of the vectors in Equation 4.10 into the matrices $\frac{\partial \pi(\underline{\beta}_x)}{\partial \underline{\beta}_{r'}}$ or $\frac{\partial \pi(\underline{\beta}_x)}{\partial \underline{\theta}_x}$. We shall discuss $\frac{\partial \int_{I_{r'j_{r'}}} f_{r'}(y_{r'} | x = x_{j_x}, \underline{\beta}_{r'}) dy_{r'}}{\partial \underline{\beta}_{r'}}$, which is part of $\frac{\partial A(\cdot)}{\partial \beta_{r'i}}$, in two cases in the next subsection where we discuss the integration required in the categorization step described earlier in this chapter. The first case will be when closed forms for the necessary integration are available and the second case when no closed forms exist and numerical integration must be utilized.

In applications, PROC NLIN and PROC IML, from the SAS software package, are used together to obtain EGNLS estimates of $\underline{\beta}$ by the modified Gauss-Newton method. Other, more “programming” intensive packages could be used to obtain these estimates via modified Gauss-Newton, such as OX or R.

4.1.3 Integration Examples of the Categorization Step

In this subsection we discuss the required integration over the continuous Y_r which is needed to calculate the conditional cell probabilities in the categorization step leading to the extended Latent Class Model in Equation 4.3. The probability of falling into region I_{rj_r} of the r th manifest variable, if that variable is continuous,

is found, in general, by the following integration:

$$\begin{aligned}
 P(Y_r \in I_{rj_r} \mid x \in I_{j_z}, \underline{\beta}_r) &\doteq \int_{l_{rj_r}} f_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) dy_r \\
 &= \int_{l_{rj_r}}^{u_{rj_r}} f_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) dy_r \\
 &= F_{Y_r|x}(u_{rj_r}) - F_{Y_r|x}(l_{rj_r}),
 \end{aligned}$$

where (l_{rj_r}, u_{rj_r}) are the endpoints of category I_{rj_r} and $F_{Y_r|x}(\cdot)$ is the cumulative distribution function for Y_r given x . For distributions with closed form CDF's, such integrals can be precisely calculated. In other circumstances, one must use numerical integration. Here we shall discuss both cases; one where all integrals have closed form solutions and one where numerical integration must be implemented. Whether this integral can be directly calculated or not, affects how the derivatives, $\frac{\partial \mathbf{A}(\cdot)}{\partial \beta_{r'i}}$, in the EGNLS step are calculated.

Case 1

We have that $P(Y_r \in I_{rj_r} \mid x \in I_{j_z}, \underline{\beta}_r) \doteq F_{Y_r|x}(u_{rj_r}) - F_{Y_r|x}(l_{rj_r})$, in general, assuming closed form solutions of the CDF's exist. In this case, the derivatives required of the right hand side of Equation 4.3 can be calculated directly since the CDF's are in closed form and in terms of the compound mappings of model parameters, i.e., $\underline{T}_r^{-1}(\underline{h}_r(x; \underline{\beta}_r))$, so one may proceed directly with EGNLS.

Suppose, for example, that we have univariate Y_r which are distributed Logistic(λ_r, p_r), for each r , where λ_r is the mean and the variance is $p_r^2 \pi^2/3$. It is known that the logistic distribution is symmetric about its mean and with only slightly heavier tails than the normal distribution. Therefore, in practice, the assumption of a normally distributed variable is essentially interchangeable with the logistic assumption. The logistic distribution has the advantage of having a closed form CDF.

For this example, we assume there are no discrete or categorical manifest variables. Then it can be shown that for each Y_r ,

$$P(Y_r \in I_{rj_r} \mid x \in I_{j_x}, \underline{\theta}_r) = \int_{l_{rj_r}}^{u_{rj_r}} (1/p_r) [\exp -(y_r - \lambda_r)/p_r] \times [1 + \exp -(y_r - \lambda_r)/p_r]^{-2} dy_r,$$

where recall (l_{rj_r}, u_{rj_r}) is the notation indicating the endpoints of interval I_{rj_r} , $j_r = 1, 2, \dots, C_r$, in the categorized version of Y_r . Assuming one was interested in modeling the mean (and, as a consequence of the GSME model specifications, the variance) for each Y_r , the mappings of Equation 3.4, defined in the beginning of Chapter 3, would be $\lambda_r = \gamma_{r1} = h_{r1}(x; \underline{\beta}_{r1})$, where $\gamma_{r1} = E(Y_r \mid x)$, and $p_r^2\pi^2/3 = \gamma_{r2} = \beta_{r2}$, where $\gamma_{r2} = Var(Y_r \mid x)$ and β_{r2} is a constant, for all r . Note that $h_{r1}(\cdot)$ is any specified, perhaps nonlinear, mean function in terms of x and $\underline{\beta}_{r1}$. Using the closed form for the CDF of the logistic distribution, we have for each r

$$P(Y_r \in I_{rj_r} \mid x \in I_{j_x}, \underline{\beta}_r) = \left[1 + \exp \frac{-\pi \left(u_{rj_r} - h_{r1}(x; \underline{\beta}_{r1}) \right)}{\sqrt{3}\beta_{r2}} \right]^{-1} - \left[1 + \exp \frac{-\pi \left(l_{rj_r} - h_{r1}(x; \underline{\beta}_{r1}) \right)}{\sqrt{3}\beta_{r2}} \right]^{-1}.$$

The probabilities for the categories of x are approximated as $P(x \in I_{j_x} \mid \underline{\theta}_x) \doteq f(x_{j_x} \mid \underline{\theta}_x)L_{j_x}$ and no integration is required, so the distribution of x need not be assumed to follow a logistic distribution, or any other distribution with a closed form CDF.

The extended Latent Class Model in Equation 4.2, for the current example is $P(Y_1 \in I_{1j_1}, Y_2 \in I_{2j_2}, \dots, Y_p \in I_{pj_p}) \doteq \sum_{j_x} \prod_{r=1}^p P(Y_r \in I_{rj_r} \mid x \in I_{j_x}, \underline{\beta}_r)P(x \in I_{j_x} \mid \underline{\theta}_x)$. Here, closed form solutions for CDF's allow the probabilities on the right hand side of the equation to be in terms of the model parameters, i.e., $\underline{\beta} = (\underline{\beta}'_1, \underline{\beta}'_2, \dots, \underline{\beta}'_p, \underline{\theta}'_x)'$. So, the elements for the left hand side of Equation 4.2, $P(Y_1 \in I_{1j_1}, Y_2 \in I_{2j_2}, \dots, Y_p \in I_{pj_p})$, are estimated by multinomial

sample proportions, i.e., $\hat{\underline{\pi}}$, calculated from the data, and the right hand side, $\sum_j \prod_{r=1}^p P(Y_r \in I_{rj_r} \mid x \in I_{j_z}, \underline{\beta}_r) P(x \in I_{j_z} \mid \underline{\theta}_x)$, are nonlinear functions of the model parameters, i.e., $\underline{\beta}$. We are now ready to apply EGNLS to estimate the GSME model parameters, $\underline{\beta}$.

EGNLS requires consistent initial estimates for favorable asymptotic properties of the estimators. We shall discuss a method for developing these consistent estimates at the end of this chapter. Application of EGNLS involves derivatives with respect to the model parameters and then using these derivatives in Equation 4.9. Since in the current example, closed form solutions of the necessary integration required to calculate probabilities for the categories of the variables exist, the required piece of $\frac{\partial \Lambda(\cdot)}{\partial \beta_{r'i}}$ for use in the EGNLS procedure is straight forward to calculate. In the current example, the T_{r1} mapping is the identity for each variable, so for the r' th variable, we have:

$$\begin{aligned} & \frac{\partial P(Y_{r'} \in I_{r'j_{r'}} \mid x \in I_{j_z}, \underline{\beta}_{r'})}{\partial h_{r'1}} = \\ & \frac{\partial}{\partial (h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))} \left(\left[1 + \exp \frac{-\pi (u_{r'j_{r'}} - h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right]^{-1} \right. \\ & \quad \left. - \left[1 + \exp \frac{-\pi (l_{r'j_{r'}} - h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right]^{-1} \right) \\ & = - \left[1 + \exp \frac{-\pi (u_{r'j_{r'}} - h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right]^{-2} \left(\exp \frac{-\pi (u_{r'j_{r'}} - h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right) \times \\ & \quad \frac{\partial}{\partial (h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))} \left(\exp \frac{-\pi (u_{r'j_{r'}} - h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right) \\ & \quad + \left[1 + \exp \frac{-\pi (l_{r'j_{r'}} - h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right]^{-2} \left(\exp \frac{-\pi (l_{r'j_{r'}} - h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right) \\ & \quad \times \frac{\partial}{\partial (h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))} \left(\exp \frac{-\pi (l_{r'j_{r'}} - h_{r'1}(x_{j_z}; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right). \end{aligned}$$

This part of $\frac{\partial A(\cdot)}{\partial \underline{\beta}_{r'1}}$ would be used when taking derivatives with respect to the elements of $\underline{\beta}_{r'1}$. EGNLS requires derivatives with respect to all the model parameters, so we also must consider the derivative with respect to the constant $\beta_{r'2}$.

The compound mappings associated with this parameter are far less complex than those for the elements $\underline{\beta}_{r'1}$ due to the fact a functional form must be specified for $h_{r'1}(x_{jz}; \underline{\beta}_{r'1})$, whereas the form for the constant variance is known. The necessary pieces of the derivative $\frac{\partial A(\cdot)}{\partial \underline{\beta}_{r'1}}$ reduce to $\frac{\partial \int_{I_{r'j_{r'}}} f_{r'}(y_{r'} | x=x_{jz}, \underline{\beta}_{r'}) dy_{r'}}{\partial \beta_{r'2}}$, where $\underline{\beta}_{r'2}$ is the scalar $\beta_{r'2}$. This derivative is

$$\begin{aligned} \frac{\partial P(Y_{r'} \in I_{r'j_{r'}} | x \in I_{jz}, \underline{\beta}_{r'})}{\partial \beta_{r'2}} &= \\ \frac{\partial}{\partial \beta_{r'2}} \left(\left[1 + \exp \frac{-\pi (u_{r'j_{r'}} - h_{r'1}(x; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right]^{-1} - \left[1 + \exp \frac{-\pi (l_{r'j_{r'}} - h_{r'1}(x; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right]^{-1} \right) \\ &= - \left[1 + \exp \frac{-\pi (u_{r'j_{r'}} - h_{r'1}(x; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right]^{-2} \exp \frac{-\pi (u_{r'j_{r'}} - h_{r'1}(x; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \\ &\quad \times \frac{-3\pi (u_{r'j_{r'}} - h_{r'1}(x; \underline{\beta}_{r'1}))}{2(3\beta_{r'2})^{3/2}} + \left[1 + \exp \frac{-\pi (l_{r'j_{r'}} - h_{r'1}(x; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \right]^{-2} \\ &\quad \times \exp \frac{-\pi (l_{r'j_{r'}} - h_{r'1}(x; \underline{\beta}_{r'1}))}{\sqrt{3\beta_{r'2}}} \times \frac{-3\pi (l_{r'j_{r'}} - h_{r'1}(x; \underline{\beta}_{r'1}))}{2(3\beta_{r'2})^{3/2}}. \end{aligned}$$

These derivatives are used in computing the estimate at iteration $k+1$, by finding $\underline{D}_k = \left[\left(\frac{\partial \pi(\underline{\beta}_k)}{\partial \underline{\beta}} \right)' \hat{V}^{-1} \left(\frac{\partial \pi(\underline{\beta}_k)}{\partial \underline{\beta}} \right) \right]^{-1} \left(\frac{\partial \pi(\underline{\beta}_k)}{\partial \underline{\beta}} \right)' \hat{V}^{-1} [\hat{\pi} - \pi(\underline{\beta}_k)]$, and then a $0 < \lambda_k < 1$ such that $Q(\hat{\underline{\beta}}_k + \lambda_k \underline{D}_k, \hat{V}) < Q(\hat{\underline{\beta}}_k, \hat{V})$.

In any application where all distributions can be integrated in a closed form, such as the one just discussed, any of a number of computer packages may be utilized to perform the necessary calculations. Examples of such packages are SAS using PROC IML, OX, and R.

Case 2

The second case for calculating the integrals needed in the translation to the Latent Class model assumes no closed form solutions exist for the CDF's and

numerical integration must therefore be implemented. This is the case that would most likely be used in practice. We would like to perform the following integration:

$$\begin{aligned}
 P(Y_r \in I_{rj_r} \mid x \in I_{j_x}, \underline{\beta}_r) &= \int_{I_{rj_r}} f_r(y_r \mid x = x_{j_x}, \underline{\beta}_r) dy_r \\
 &= \int_{l_{rj_r}}^{u_{rj_r}} f_r(y_r \mid x = x_{j_x}, \underline{\beta}_r) dy_r \\
 &= F_{Y_r|x}(u_{rj_r}) - F_{Y_r|x}(l_{rj_r}),
 \end{aligned}$$

however, no closed forms exist for $F_{Y_r|x}(\cdot)$. Using EGNLS to estimate the parameters of Equation 4.2, will require derivatives with respect to $\underline{\beta}_r$. For this case, the appropriate derivative for the r' th variable to be used in $\frac{\partial A(\cdot)}{\partial \beta_{r'i}}$ in the EGNLS procedure is:

$$\frac{\partial \int_{I_{r'j_{r'}}} f_{r'}(y_{r'} \mid x = x_{j_x}, \underline{\beta}_{r'}) dy_{r'}}{\partial \underline{T}_{r'}^{-1}} = \int_{I_{r'j_{r'}}} \frac{\partial f_{r'}(y_{r'} \mid x = x_{j_x}, \underline{\beta}_{r'})}{\partial \underline{T}_{r'}^{-1}} dy_{r'},$$

assuming the order of differentiation and integration is interchangeable. If we know, for example, that we are within the exponential family, then this interchangeability is guaranteed. In applications, since the distribution for $Y_{r'}$ is known, the derivative inside the integral can be computed and then initial starting values are used to perform numerical integration.

Let us assume that $Y_{r'}$ is a univariate, continuous variable in the single parameter exponential family. Then for the right hand side of the previous equation, with $T_{r'}^{-1}(h_{r'}(x; \underline{\beta}_{r'})) = T_{r'}^{-1}$ in place of the canonical parameter,

$\theta_{r'}$, we would have

$$\begin{aligned}
 & \int_{I_{r',j_{r'}}} \frac{\partial f_{r'}(y_{r'} \mid x = x_{j_x}, \underline{\beta}_{r'})}{\partial T_{r'}^{-1}} dy_{r'} \\
 &= \int_{I_{r',j_{r'}}} \frac{\partial \exp\{[y_{r'} T_{r'}^{-1} - b_{r'}(T_{r'}^{-1})]/a_{r'}(\phi_{r'}) + c_{r'}(y_{r'}, \phi_{r'})\}}{\partial T_{r'}^{-1}} dy_{r'} \\
 &= \int_{I_{r',j_{r'}}} [\exp\{[y_{r'} T_{r'}^{-1} - b_{r'}(T_{r'}^{-1})]/a_{r'}(\phi_{r'}) + c_{r'}(y_{r'}, \phi_{r'})\}] \\
 &\quad \times [y_{r'} - b'_{r'}(T_{r'}^{-1})]/a_{r'}(\phi_{r'}) dy_{r'},
 \end{aligned}$$

where $b'_{r'}(T_{r'}^{-1})$ is the derivative of $b_{r'}(T_{r'}^{-1})$ with respect to $T_{r'}^{-1}$. This integral, with specified starting values, can now be calculated using numerical integration and thus be used in the appropriate place in calculating $\frac{\partial A(\cdot)}{\partial \beta_{r'i}}$. So, in such a case where there are no closed forms for the CDF's, the derivatives required in the EGNLS procedure are pushed through the integration first, and then after the derivatives are taken and estimates substituted for $\underline{\beta}$, numerical integration is utilized. In applications, one of the many mathematical-type software packages that can perform numerical integration can be used. MATLAB, MAPLE, and MATHEMATICA are three such software packages that could be used.

4.2 Generalization to Estimating a Multivariate Model

Now that we have seen the procedure for estimation and the notation that is required in the simple case of all univariate manifest variables, we shall show by example how the grouping of a set of continuous variables that are conditionally dependent with each other is handled. Together, these continuous variables are conditionally independent, given x , of the remaining manifest variables. We shall illustrate such grouping into conditionally independent subsets of variables in the context of the multivariate simple linear regression model with measurement error in x . Let the multivariate response vector be the outcome vector of main interest, $\underline{Y}_r$, in the GSME model. This is an important application because it allows the GSME model to be applied to multivariate regression situations, as discussed in

Section 3.2 of Chapter 3, where we saw the numerous applications of the GSME model. Recall from that section in Chapter 3, that a nonlinear multivariate regression model for a vector valued variable outcome of main interest would be:

$$\begin{bmatrix} Y_{r^*1} \\ Y_{r^*2} \\ \vdots \\ Y_{r^*p_{r^*}} \end{bmatrix} = \begin{bmatrix} \mu_{r^*1}(x; \underline{\beta}_{r^*1}) \\ \mu_{r^*2}(x; \underline{\beta}_{r^*2}) \\ \vdots \\ \mu_{r^*p_{r^*}}(x; \underline{\beta}_{r^*p_{r^*}}) \end{bmatrix} + \underline{e}_{r^*},$$

where $\mu_{r^*s}(\cdot)$ is an arbitrary nonlinear mean function for the s th variable, $s = 1, 2, \dots, p_{r^*}$, and we assume $E(\underline{e}_{r^*}) = \underline{0}$ and $Var(\underline{e}_{r^*}) = \Sigma_{e_{r^*}}$. We know that $\Sigma_{e_{r^*}}$ is not a diagonal matrix due to the conditional dependence of the variables which make up the multivariate response variable. Assume the remaining manifest variables are univariate and that they are conditionally independent.

We know that there are p_{r^*} continuous variables which are conditionally dependent and that they are grouped together to form a multivariate manifest vector $\underline{Y}_{r^*}$. The joint conditional distribution of the continuous variables is known and given by $f_{r^*}(\underline{y}_{r^*} | x, \underline{\theta}_{r^*})$. Recall that the dimension of $\underline{Y}_{r^*}$ is p_{r^*} and that of $\underline{\theta}_{r^*}$ is k_{r^*} . The mappings between the conditional distribution parameters and the model parameters, $\underline{\beta}_{r^*}$, is $\underline{\theta}_{r^*} = \underline{T}_{r^*}^{-1}(\underline{h}_{r^*}(x; \underline{\beta}_{r^*}))$, which were defined in Equation 3.4, and still hold. The dimension of $\underline{\beta}_{r^*}$ is l_{r^*} .

For $\underline{Y}_{r^*}$ given x we have $\underline{Y}_{r^*} = (Y_{r^*1}, Y_{r^*2}, \dots, Y_{r^*p_{r^*}})'$, where each univariate variable which makes up $\underline{Y}_{r^*}$ is categorized just as those described in the all univariate case. The number of categories of each variable may be given in vector notation as $\underline{C}_{r^*} = (C_{r^*1}, C_{r^*2}, \dots, C_{r^*p_{r^*}})'$. For each Y_{r^*s} define *intervals* $(l_{r^*sj_{r^*s}}, u_{r^*sj_{r^*s}})$, $s = 1, 2, \dots, p_{r^*}$, and $j_{r^*s} = 1, 2, \dots, C_{r^*s}$. Denote the interval defining the j_{r^*s} th category for variable s of $\underline{Y}_{r^*}$ as $I_{r^*sj_{r^*s}}$ and the length($I_{r^*sj_{r^*s}}$) by $L_{r^*sj_{r^*s}}$. For each $\underline{Y}_{r^*}$, the intersection of these intervals define a *region* which can be thought of as a multidimensional interval. The region defining

the $j_{r^*1}j_{r^*2}\dots j_{r^*p_{r^*}}$ th multidimensional interval is

$$R_{r^*j_{r^*}} = \left[(l_{r^*1j_{r^*1}}, u_{r^*1j_{r^*1}}) \times (l_{r^*2j_{r^*2}}, u_{r^*2j_{r^*2}}) \times \dots \times (l_{r^*p_{r^*}j_{r^*p_{r^*}}}, u_{r^*p_{r^*}j_{r^*p_{r^*}}}) \right]$$

where $j_{r^*} = j_{r^*1}j_{r^*2}\dots j_{r^*p_{r^*}}$ and “ $\times$ ” is defined to be the *Cartesian product* as found, for example, in Khuri [34, p. 3]. Alternatively, we may write $R_{r^*j_{r^*}} = \times_{s=1}^{p_{r^*}} (l_{r^*sj_{r^*s}}, u_{r^*sj_{r^*s}})$. To indicate the probability that the outcome manifest vector of interest falls into a certain region, we write $P(Y_{r^*} \in R_{r^*j_{r^*}})$. Note, again, that the intervals need not be of equal length, either between or within the sets of continuous manifest variables. Calculating the cell probabilities for Y_{r^*} is done in a similar manner as that described earlier for a univariate continuous variable accept multivariate integration is performed. We have

$$\begin{aligned} P(Y_{r^*} \in R_{r^*j_{r^*}} \mid x \in I_{j_x}, \underline{\beta}_{r^*}) &= \frac{P(Y_{r^*} \in R_{r^*j_{r^*}}, x \in I_{j_x} \mid \underline{\beta}_{r^*})}{P(x \in I_{j_x} \mid \underline{\theta}_x)} \\ &= \frac{\int_{R_{r^*j_{r^*}}} \int_{I_{j_x}} f_{r^*}(y_{r^*}, x \mid \underline{\theta}_{r^*x}) dx dy_{r^*}}{\int_{I_{j_x}} f(x \mid \underline{\theta}_x) dx} \\ &= \frac{\int_{R_{r^*j_{r^*}}} f_{r^*}(y_{r^*}, x_{j_x} \mid \underline{\theta}_{r^*x}) L_{j_x} dy_{r^*}}{f(x_{j_x} \mid \underline{\theta}_x) L_{j_x}} \\ &= \int_{R_{r^*j_{r^*}}} f_{r^*}(y_{r^*} \mid x = x_{j_x}, \underline{\beta}_{r^*}) dy_{r^*}, \end{aligned}$$

for I_{j_x} of small length, dy_{r^*} corresponding to the multidimensional integration over $R_{r^*j_{r^*}}$, $\int_{R_{r^*j_{r^*}}} = \int_{I_{r^*1j_{r^*1}}} \int_{I_{r^*2j_{r^*2}}} \dots \int_{I_{r^*p_{r^*}j_{r^*p_{r^*}}}$, and, where $\underline{\theta}_{r^*x}$ are the parameters of the joint distribution of Y_{r^*} and x , which are unknown. Again, it is in this integration step where the substitution for the conditional distribution parameters, defined by $\underline{\theta}_{r^*} = \underline{T}_{r^*}^{-1}(\underline{h}_{r^*}(x; \underline{\beta}_{r^*}))$, take place to allow the model parameters, $\underline{\beta}_{r^*}$, to enter the nonlinear functions defining the cell probabilities as function of $\underline{\beta}_{r^*}$.

For the remaining $p - 1$ univariate manifest variables (minus one because the outcome of main interest is accounted for by r^*) order them, without loss of

generality, so that Y_r , $r = 1, 2, \dots, p'$, are the continuous variables, Y_r , $r = p' + 1, p' + 2, \dots, p''$, are the “other” discrete variables, and Y_r , $r = p'' + 1, p'' + 2, \dots, p - 1$, are the multinomial categorical variables. Those remaining manifest variables which require categorization are categorized just as described above, for the all univariate case. The goal in the estimation procedure is still to arrive at a Latent Class Model version of the joint marginal distribution of the manifest variables, given in Equation 4.2. After arriving at such an analog equation, we apply EGNLS to estimate the parameters of the model $\hat{\pi} = \pi(\underline{\beta}) + \varepsilon$, i.e., Equation 4.4. We have as the probability of falling into cell $j_{r^*} j_1 j_2 \dots j_{p-1}$ of the multinomial distribution:

$$\begin{aligned} \pi_{j_{r^*} j_1 j_2 \dots j_{p-1}}(\underline{\beta}) &\doteq \sum_{j_z} \left[\int_{R_{r^*} j_{r^*}} f_{r^*}(y_{r^*} \mid x = x_{j_z}, \underline{\beta}_{r^*}) dy_{r^*} \right] \\ &\prod_{r=1}^{p'} \left[\int_{I_{r j_r}} f_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) dy_r \right] \\ &\prod_{r=p'+1}^{p''} \left[\sum_{y_r \in I_{r j_r}} p_r(y_r \mid x = x_{j_z}, \underline{\beta}_r) \right] \\ &\prod_{r=p''+1}^{p-1} P(Y_r = j_r \mid x = x_{j_z}, \underline{\beta}_{r^*}) f(x_{j_z} \mid \underline{\theta}_x) L_{j_z}, \end{aligned}$$

where $I(r \leq p')$ and $I(p' + 1 \leq r \leq p'')$ are indicator variables. As in the univariate case, these probabilities are concatenated together and used, along with the sample proportions, to form the model $\hat{\pi} = \pi(\underline{\beta}) + \varepsilon$, which is the starting point of the EGNLS procedure. The dimension of $\pi(\underline{\beta})$ is now $N = (C_{r^*1} \times C_{r^*2} \times \dots \times C_{r^*p^*}) \prod_{r=1}^{p-1} C_r$. Recall, however, that the last element is dropped to avoid the singularity of V in the estimation process. The sample proportions, $\hat{\pi}$, are maximum likelihood estimators for the vector of the multinomial cell probabilities given by, in this example, $P(\underline{Y}_{r^*} \in R_{r^*} j_{r^*}, Y_1 \in I_{1 j_1}, \dots, Y_{p''} \in$

$I_{p''j_{p''}}, Y_{p''+1} = j_{p''+1}, \dots, Y_{p-1} = j_{p-1}$), for all possible combinations of cross-classifications. The estimation procedure is essentially the same as the univariate case, from this point.

Note that the univariate methods in this chapter can be extended to the general multivariate case, where each manifest response variable may be a vector. The notation, however, is messy when the elements of $\underline{Y}_r$ are allowed to be any mix of continuous, discrete, and categorical variables.

4.3 Asymptotic Properties

In this section, we examine the asymptotic properties and, therefore, develop a means of inference for the estimators derived for the GSME model. The following results hold for both examples previously given, i.e., the all univariate case and the multivariate regression case, as well as for the model in its full generality. This is because the theorems and their proofs do not depend on specifications of the different models. Only minor details which have no effect on the results, such as the value of N (which is the dimension of $\underline{\pi}$), would change depending on the model.

Recall that the main idea is to estimate the model parameters, $\underline{\beta}$, by minimizing the quadratic form, $Q(\underline{\beta}, \hat{V})$, given in Equation 4.5. This is done by setting $\frac{\partial}{\partial \underline{\beta}'} Q(\underline{\beta}, \hat{V})$ equal to zero and solving for $\underline{\beta}$, which is equivalent to solving the estimating equations given in Equation 4.7 for $\underline{\beta}$, i.e., finding the solution to

$$\underline{S}_n(\underline{\beta}) \equiv \left(\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'} \right)' \hat{V}^{-1} (\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta})) = \underline{0},$$

where $\hat{V}$ is a positive definite matrix which is consistent for V . Recall we established these estimating equations by letting $\underline{S}_n(\underline{\beta}) = \left(\frac{\partial}{\partial \underline{\beta}'} Q(\underline{\beta}, \hat{V}) \right)' / (-2n)$.

The method of EGNLS requires using consistent initial estimates to obtain favorable asymptotic properties. In the next section we discuss deriving such consistent estimates for use in the initial step of the EGNLS procedure. It is worth

noting that $\underline{\pi}(\underline{\beta})$ does not depend on x , or rather x_{j_x} , the midpoints of the intervals after categorization, since in deriving the joint distribution of the categorical $\underline{Y}_1, \underline{Y}_2, \dots, \underline{Y}_p$, summation over all x_{j_x} it performed.

In order to arrive at a lemma which will be used in developing the asymptotic properties of the EGNLS estimators, we must consider the derivative, with respect to $\underline{\beta}'$, of the left hand side of Equation 4.7, which is proportionately equivalent to the transpose of the second derivative of the quadratic form $Q(\underline{\beta}, \hat{V})$ with respect to $\underline{\beta}$, given earlier in this chapter. Initially, we must differentiate $S_n(\underline{\beta})$ with respect to the individual elements of the vector of parameters $\underline{\beta}$. For ease of notation, in this section only, we shall define these individual elements as β_i , $i = 1, 2, \dots, \nu$, where, again, the last k_x of the β_i are representations of the elements of $\underline{\theta}_x$. The result which was established earlier in this chapter for the derivative of multiply compounded, multivariable functions will be utilized throughout this section. Recall the $(N - 1) \times \nu$ matrix $\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'}$ is defined using that result, giving as its (j, i) th element

$$\left(\frac{\partial}{\partial \underline{T}^{-1}} \pi_j(\underline{T}^{-1}) \right) \left(\frac{\partial}{\partial \underline{h}} \underline{T}^{-1}(\underline{h}) \right) \left(\frac{\partial}{\partial \beta_i} f_1(\underline{x}) \right).$$

Differentiating the left hand side of Equation 4.7 with respect to β_i , i.e.,

$$\frac{\partial}{\partial \beta_i} S_n(\underline{\beta}) = \frac{\partial}{\partial \beta_i} \left[\left(\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'} \right)' \hat{V}^{-1} (\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta})) \right],$$

will result in column vectors which when concatenated side by side form the $\nu \times \nu$ matrix $\frac{\partial}{\partial \underline{\beta}'} S_n(\underline{\beta})$. Applying the product rule for matrices [49, p. 329] gives

$$\begin{aligned} \frac{\partial}{\partial \beta_i} S_n(\underline{\beta}) &= \left(\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'} \right)' \left[\frac{\partial}{\partial \beta_i} \hat{V}^{-1} (\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta})) \right] + \\ &\quad \left[\frac{\partial}{\partial \beta_i} \left(\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'} \right)' \right] \hat{V}^{-1} (\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta})). \end{aligned} \quad (4.11)$$

Upon concatenation of all the resulting column vectors from the first piece on the right hand side of Equation 4.11 to form the columns of a matrix, and similarly, concatenating the resulting vectors from the second piece on the right hand side of Equation 4.11 into one matrix, we have as the derivative to the left side of Equation 4.7:

$$\frac{\partial}{\partial \underline{\beta}'} \underline{S}_n(\underline{\beta}) = \underline{S}'_n(\underline{\beta}) = -\underline{U}_n(\underline{\beta}) + \underline{R}_n(\underline{\beta}),$$

where

$$\underline{U}_n(\underline{\beta}) = \left(\frac{\partial \underline{\pi}(\underline{\beta})'}{\partial \underline{\beta}} \right) \hat{V}^{-1} \left(\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'} \right),$$

the (a, b) th element of $\underline{R}_n(\underline{\beta})$ is

$$r_{ab} = \sum_i \sum_j (\hat{\pi}_j - \pi_j(\underline{\beta})) \left(\frac{\partial^2 \pi_i(\underline{\beta})}{\partial \beta_a \partial \beta_b} \hat{v}^{ij} \right),$$

and $\pi_j(\underline{\beta})$ and $\hat{\pi}_j$ are, respectively, the j th elements of the vectors $\underline{\pi}(\underline{\beta})$ and $\hat{\underline{\pi}}$.

Further, $\hat{v}^{ij}$ is the (i, j) th element of the matrix $\hat{V}^{-1}$, and β_a and β_b are the a th and b th elements of $\underline{\beta}$.

Now we must introduce some notation and definitions. Following Foutz [21], define $\|M\|$ to be the *norm* of an $r \times r$ dimensional matrix M which is the least upper bound (lub) of all numbers $|M\underline{z}|$, where $\underline{z}$ ranges over all vectors in Euclidean r space, E^r , such that $|\underline{z}| \leq 1$, and define $|\underline{z}|$ to be the *Euclidean norm*, or length, of a vector which is equal to $(\underline{z}'\underline{z})^{1/2} = (\sum_{i=1}^r z_i^2)^{1/2}$. Further, if $\underline{z}$ is a scalar, then $|\underline{z}|$ is the *absolute value* of $\underline{z}$. It is worth noting for future use that the norm, Euclidean norm, and absolute value have similar properties. For example, the triangle inequality holds for all three (see [21, 49, p.158]). Throughout this section, we also make use of many known large sample theory results. For example, the concepts of order in probability, $o_p(\cdot)$, and bounded in probability, $O_p(\cdot)$, are used and known properties of these large sample concepts may be applied without restating the specific result. As a specific example, later in this section we use the

property that if $A_n = O_p(f_n)$ and $B_n = O_p(g_n)$, then $A_n + B_n = O_p(\max(f_n, g_n))$, without specifically stating this large sample result prior to its use. An informative, general summary of these concepts and their results may be found in Chapter 5 of Fuller [22]. Agresti [1] provides a review of some large sample results specifically for multinomial data in Chapter 12 of his book on categorical data analysis.

We now state and prove a lemma used to establish the asymptotic properties of the EGNLS estimator for the GSME model.

LEMMA 2: Let $\underline{\beta}^0$ and $\underline{\pi}^0$ denote the true values of $\underline{\beta}$ and $\underline{\pi}(\underline{\beta})$. Assume that in an open neighborhood of $\underline{\beta}^0$, defined by $D_\delta \equiv \{\underline{\beta} : |\underline{\beta} - \underline{\beta}^0| < \delta\}$, the following regularity conditions hold, for all $\underline{\beta} \in D_\delta$:

1. $\pi_j(\underline{\beta}) > 0$, $j = 1, 2, \dots, N$, where the calculation of N may vary in different applications. For example, we saw in the univariate case that $N = \prod_{r=1}^p C_r$, whereas in the more general case shown for multivariate regression of the outcome vector of interest, $N = (C_{r+1} \times C_{r+2} \times \dots \times C_{r+p_{r,\bullet}}) \prod_{r=1}^{p-1} C_r$,
2. the second order partial derivatives of $\underline{\pi}(\underline{\beta})$ exist and are continuous,
3. the Jacobian matrix $(\partial \underline{\pi}(\underline{\beta}) / \partial \underline{\beta}')$ has full column rank $\nu = (l_1 + l_2 + \dots + l_p + k_x)$, where, recall, l_r is the dimension of $\underline{\beta}_r$, for each r , and k_x is the dimension of $\underline{\theta}_x$.

Then

$$(A) \quad \underline{S}_n(\underline{\beta}^0) \xrightarrow{P_{\underline{\beta}^0}} \underline{0}, \text{ as } n \rightarrow \infty,$$

$$(B) \quad \underline{U}_n(\underline{\beta}^0) \xrightarrow{P_{\underline{\beta}^0}} \underline{U}(\underline{\beta}^0) \text{ and } \underline{U}(\underline{\beta}) \text{ is positive definite for all } \underline{\beta} \in D_\delta \text{ and is}$$

continuous at $\underline{\beta}^0$, and

$$P_{\underline{\beta}^0}$$

(C) $\sup_{\underline{\beta} \in D_\delta} \|S'_n(\underline{\beta}) + U(\underline{\beta})\| \longrightarrow 0$, as $n \rightarrow \infty$;

Proof: (A) We have that $\underline{S}_n(\underline{\beta}^0) = \left(\frac{\partial \underline{\pi}(\underline{\beta}^0)}{\partial \underline{\beta}^{0'}} \right)' \hat{V}^{-1} (\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta}^0))$, where it is known that

$$P_{\underline{\beta}^0}$$

$$\hat{\underline{\pi}} \longrightarrow \underline{\pi}(\underline{\beta}^0),$$

as $n \rightarrow \infty$. Further, due to regularity condition 2, the elements of $\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'}$ exist and are continuous and bounded in D_δ , therefore the elements of $\frac{\partial \underline{\pi}(\underline{\beta}^0)}{\partial \underline{\beta}^{0'}}$ are bounded. Moreover, $\hat{V}^{-1}$ converges to V^{-1} in probability, as $n \rightarrow \infty$, because $\hat{V}$ converges in probability to V and the elements of V^{-1} are continuous functions of the elements of V . Since V^{-1} is a positive definite matrix, $\hat{V}^{-1}$ must be bounded, with probability going to one, as $n \rightarrow \infty$, for all n greater than some sufficiently large

$$P_{\underline{\beta}^0}$$

N_0 . Therefore, $\underline{S}_n(\underline{\beta}^0) \longrightarrow \underline{0}$, as $n \rightarrow \infty$.

(B) We have that

$$U_n(\underline{\beta}^0) = \left(\frac{\partial \underline{\pi}(\underline{\beta}^0)}{\partial \underline{\beta}^{0'}} \right)' \hat{V}^{-1} \left(\frac{\partial \underline{\pi}(\underline{\beta}^0)}{\partial \underline{\beta}^{0'}} \right),$$

where $\frac{\partial \underline{\pi}(\underline{\beta}^0)}{\partial \underline{\beta}^{0'}}$ does not depend on n and has elements which are bounded (by condition 2) and $\hat{V}^{-1}$ converges to V^{-1} in probability, as $n \rightarrow \infty$, as shown in the proof of (A). Therefore

$$P_{\underline{\beta}^0}$$

$$U_n(\underline{\beta}^0) \longrightarrow U(\underline{\beta}^0),$$

where

$$U(\underline{\beta}^0) = \left(\frac{\partial \pi(\underline{\beta}^0)}{\partial \underline{\beta}^{0'}} \right)' V^{-1} \left(\frac{\partial \pi(\underline{\beta}^0)}{\partial \underline{\beta}^{0'}} \right).$$

Further, we know that V^{-1} is positive definite, thus, for all $\underline{z} \neq 0$,

$$\underline{z}' [U(\underline{\beta})] \underline{z} = \underline{z}' \left[\left(\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'} \right)' V^{-1} \left(\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'} \right) \right] \underline{z} > 0.$$

Therefore, $U(\underline{\beta})$ is positive definite for all $\underline{\beta}$ in a neighborhood of $\underline{\beta}^0$, i.e., for all $\underline{\beta} \in D_\delta$.

We know $\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'}$ does not depend on n and, by condition 2, it exists and is continuous for all $\underline{\beta} \in D_\delta$, therefore it must be continuous at $\underline{\beta}^0$, where obviously $\underline{\beta}^0 \in D_\delta$. Furthermore, since V^{-1} does not depend on $\underline{\beta}^0$, we have that $U(\underline{\beta})$ is continuous at $\underline{\beta}^0$.

(C) We know that $S'_n(\underline{\beta}) = -U_n(\underline{\beta}) + R_n(\underline{\beta})$ so we may write

$$S'_n(\underline{\beta}) + U(\underline{\beta}) = -U_n(\underline{\beta}) + U(\underline{\beta}) + R_n(\underline{\beta})$$

where $-U_n(\underline{\beta}) + U(\underline{\beta}) = o_p(1)$ for all $\underline{\beta} \in D_\delta \equiv \{\underline{\beta} : |\underline{\beta} - \underline{\beta}^0| < \delta\}$, because

$$\begin{aligned} -U_n(\underline{\beta}) + U(\underline{\beta}) &= - \left(\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'} \right)' \widehat{V}^{-1} \left(\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'} \right) + \left(\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'} \right)' V^{-1} \left(\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'} \right) \\ &= \left(\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'} \right)' [-\widehat{V}^{-1} + V^{-1}] \left(\frac{\partial \pi(\underline{\beta})}{\partial \underline{\beta}'} \right), \end{aligned}$$

and $-\widehat{V}^{-1} + V^{-1} = o_p(1)$, as shown in the proof of (A). Thus $-U_n(\underline{\beta}) + U(\underline{\beta}) = o_p(1)$, for all $\underline{\beta} \in D_\delta$, and we may conclude that $\sup_{\underline{\beta} \in D_\delta} \|-U_n(\underline{\beta}) + U(\underline{\beta})\| = o_p(1)$ because the norm is a continuous function of the elements of $-U_n(\underline{\beta}) + U(\underline{\beta})$ and the norm of $-U_n(\underline{\beta}) + U(\underline{\beta})$ being $o_p(1)$ for all $\underline{\beta} \in D_\delta$, implies the norm must have the same order for the sup over all $\underline{\beta} \in D_\delta$.

Therefore, the probability limit of $\sup_{\underline{\beta} \in D_\delta} \|S'_n(\underline{\beta}) + U(\underline{\beta})\|$ is zero if the probability limit of $\sup_{\underline{\beta} \in D_\delta} \|R_n(\underline{\beta})\|$ is zero because

$$\begin{aligned} \sup_{\underline{\beta} \in D_\delta} \|S'_n(\underline{\beta}) + U(\underline{\beta})\| &= \sup_{\underline{\beta} \in D_\delta} \|-U_n(\underline{\beta}) + U(\underline{\beta}) + R_n(\underline{\beta})\| \\ &\leq \sup_{\underline{\beta} \in D_\delta} \|-U_n(\underline{\beta}) + U(\underline{\beta})\| + \sup_{\underline{\beta} \in D_\delta} \|R_n(\underline{\beta})\| \\ &= o_p(1) + \sup_{\underline{\beta} \in D_\delta} \|R_n(\underline{\beta})\|. \end{aligned}$$

$$P_{\underline{\beta}^0}$$

Hence, it remains to show that $\sup_{\underline{\beta} \in D_\delta} \|R_n(\underline{\beta})\| \rightarrow 0$, as $n \rightarrow \infty$, in

order to prove (C). Recall that one element of $R_n(\underline{\beta})$ takes the form $\sum_i \sum_j (\hat{\pi}_j -$

$$P_{\underline{\beta}^0} \pi_j(\underline{\beta})) \left(\frac{\partial^2 \pi_i(\underline{\beta})}{\partial \beta_a \partial \beta_b} \hat{v}^{ij} \right). \text{ We know that } \hat{\pi}_j \rightarrow \pi_j(\underline{\beta}^0), \text{ as } n \rightarrow \infty, \text{ for all } j. \text{ Thus for any}$$

positive δ' and ε there exists an N_0 such that for all $n \geq N_0$,

$$P(|\hat{\pi}_j - \pi_j(\underline{\beta}^0)| > \varepsilon/2) < \delta'. \quad (4.12)$$

Let δ' be such that $|\pi_j(\underline{\beta}) - \pi_j(\underline{\beta}^0)| < \varepsilon/2$ when $|\underline{\beta} - \underline{\beta}^0| < \delta'$. We know such a δ' exists for any $\varepsilon > 0$ because $\pi(\underline{\beta})$ is continuous on D_δ . Then by Equation 4.12, for all $n \geq N_0$,

$$\begin{aligned} &P(|\hat{\pi}_j - \pi_j(\underline{\beta}^0)| > \varepsilon - |\pi_j(\underline{\beta}) - \pi_j(\underline{\beta}^0)|) < \delta' \\ \Rightarrow &P(|\hat{\pi}_j - \pi_j(\underline{\beta}^0)| + |\pi_j(\underline{\beta}^0) - \pi_j(\underline{\beta})| > \varepsilon) < \delta' \\ \Rightarrow &P(|(\hat{\pi}_j - \pi_j(\underline{\beta}^0)) + (\pi_j(\underline{\beta}^0) - \pi_j(\underline{\beta}))| > \varepsilon) < \delta' \\ \Rightarrow &P(|\hat{\pi}_j - \pi_j(\underline{\beta})| > \varepsilon) < \delta'. \end{aligned}$$

So, for each element of $R_n(\underline{\beta})$, $\sum_i \sum_j (\hat{\pi}_j - \pi_j(\underline{\beta})) \left(\frac{\partial^2 \pi_i(\underline{\beta})}{\partial \beta_a \partial \beta_b} \hat{v}^{ij} \right)$, we have that $|\hat{\pi}_j - \pi_j(\underline{\beta})|$ converges in probability to zero, $\hat{v}^{ij}$ is an elements of $\hat{V}^{-1}$ which was shown in the proof of (A) to be bounded (with probability going to one as

$n \rightarrow \infty$), and by condition 2 $\frac{\partial^2 \pi_i(\underline{\beta})}{\partial \beta_a \partial \beta_b}$ are bounded for all i and $\underline{\beta} \in D_\delta$. Therefore,

$$P_{\underline{\beta}^0} \|R_n(\underline{\beta})\| \longrightarrow 0, \text{ for all } \underline{\beta} \in D_\delta, \text{ and thus the sup over all } \underline{\beta} \in D_\delta \text{ must converge in}$$

probability to zero. This completes the proof. $\diamond$

Theorem 3, and its proof, are based on similar ones found in Shah [52, p. 39] for situations involving diagnostic tests, but adapted here, for the GSME model. The theorem states the existence and, of greater importance for use with the theorems that follow, the uniqueness of a weakly consistent sequence of solutions to Equation 4.7 with limiting probability one. The key to the uniqueness of the consistent estimator is that any other consistent estimator that solves Equation 4.7 is equal to it, with probability converging to one, and therefore must have the same asymptotic distribution. The proof utilizes a version of the Inverse Function Theorem ([48, p. 221]; a simplistic interpretation of this version of the Inverse Function Theorem is that the inverse of a function exists provided its derivative has an inverse which exists, i.e., is positive or negative definite, at some point within the chosen neighborhood). The theorem and proof by Shah for the diagnostic testing situation are based on an original theorem and proof by Foutz [21] concerning unique consistent solutions to likelihood equations.

THEOREM 3: Given the regularity conditions of Lemma 2, with probability going to one as $n \rightarrow \infty$, there exists a sequence $\{\widehat{\underline{\beta}}_n\}$ of solutions to Equation 4.7, such that

$$\widehat{\underline{\beta}}_n \xrightarrow{P_{\underline{\beta}^0}} \underline{\beta}^0$$

as $n \rightarrow \infty$. Further, if $\{\bar{\underline{\beta}}_n\}$ is any other consistent sequence of solutions to Equation 4.7, then

$$P_{\bar{\underline{\beta}}^0}(\bar{\underline{\beta}}_n \neq \hat{\underline{\beta}}_n) \rightarrow 0$$

as $n \rightarrow \infty$.

Proof: By result (B) of Lemma 2, we know $U(\underline{\beta}^0)$ is positive definite. Therefore, we may define $\lambda = 1/(4\|U(\underline{\beta}^0)^{-1}\|)$. Further, we know that $S'_n(\underline{\beta}^0)$ is arbitrarily close to a negative definite matrix. This can be seen easily since $S'_n(\underline{\beta}^0) = -U_n(\underline{\beta}^0) + R_n(\underline{\beta}^0)$, where by result (B) of Lemma 2, $U_n(\underline{\beta}^0)$ converges in probability to $U(\underline{\beta}^0)$ which is positive definite and $R_n(\underline{\beta}^0) = O_p(n^{-1/2})$ because $\hat{\pi}_j - \pi_j(\underline{\beta}^0) = O_p(n^{-1/2})$, for all j , $\frac{\partial^2 \pi_i(\underline{\beta})}{\partial \beta_a \partial \beta_b}$ exists and is continuous by regularity condition 2, and $\hat{v}^{ij}$ is finite, for all i, j , because it was shown in the proof of Lemma 2 that the elements of $\hat{V}^{-1}$ are bounded. Therefore, define $\lambda_n = 1/(4\|S'_n(\underline{\beta}^0)^{-1}\|)$, whenever $S'_n(\underline{\beta}^0)$ is negative definite. Using result (C) of Lemma 2, then, we have $\lambda_n \rightarrow \lambda$ in probability.

Also by result (B), we have that $U(\underline{\beta})$ is continuous at $\underline{\beta}^0$ and one can then choose a $\delta > 0$ sufficiently small so that

$$\|U(\underline{\beta}) - U(\underline{\beta}^0)\| < \lambda/2$$

and, by part (C) of Lemma 2,

$$\lim_{n \rightarrow \infty} P_{\underline{\beta}^0} \left(\sup_{\underline{\beta} \in D_\delta} \|S'_n(\underline{\beta}) + U(\underline{\beta})\| > \lambda/4 \right) = 0$$

whenever $|\underline{\beta} - \underline{\beta}^0| < \delta$.

In the following steps, $S'_n(\underline{\beta})$ is shown to be arbitrarily close to $S'_n(\underline{\beta}^0)$ which was shown above to be arbitrarily close to a negative definite matrix. This will then allow us to apply the Inverse Function Theorem, with probability tending to

one. So, for $\underline{\beta} \in D_\delta$, i.e., $|\underline{\beta} - \underline{\beta}^0| < \delta$, we have

$$\begin{aligned}
 \|S'_n(\underline{\beta}) - S'_n(\underline{\beta}^0)\| &= \|S'_n(\underline{\beta}) + U(\underline{\beta}) - U(\underline{\beta}) - U(\underline{\beta}^0) + U(\underline{\beta}^0) - S'_n(\underline{\beta}^0)\| \\
 &\leq \|S'_n(\underline{\beta}) + U(\underline{\beta})\| + \|-S'_n(\underline{\beta}^0) - U(\underline{\beta}^0)\| + \|-U(\underline{\beta}) + U(\underline{\beta}^0)\| \\
 &< (\lambda/4) + (\lambda/4) + (\lambda/2) \\
 &= \lambda,
 \end{aligned}$$

with probability going to one as $n \rightarrow \infty$. Further, because $\lambda_n \rightarrow \lambda$ in probability, we know

$$\begin{aligned}
 P_{\underline{\beta}^0}(\|S'_n(\underline{\beta}) - S'_n(\underline{\beta}^0)\| \geq 2\lambda_n) &= P_{\underline{\beta}^0}\left(\|S'_n(\underline{\beta}) - S'_n(\underline{\beta}^0)\| \geq 2\lambda_n, \lambda_n > \frac{\lambda}{2}\right) \\
 &\quad + P_{\underline{\beta}^0}\left(\|S'_n(\underline{\beta}) - S'_n(\underline{\beta}^0)\| \geq 2\lambda_n, \lambda_n \leq \frac{\lambda}{2}\right) \\
 &\leq P_{\underline{\beta}^0}(\|S'_n(\underline{\beta}) - S'_n(\underline{\beta}^0)\| \geq 2\lambda) + P_{\underline{\beta}^0}\left(\lambda_n \leq \frac{\lambda}{2}\right) \\
 &\rightarrow 0 + 0 = 0,
 \end{aligned}$$

as $n \rightarrow \infty$.

Now we may apply the appropriate version of the Inverse Function Theorem [48, p. 221] to insure, with probability approaching one, that $\underline{S}_n$ is a one-to-one function from D_δ onto $\underline{S}_n(D_\delta)$ and $\underline{S}_n(D_\delta)$, the image set, contains the open neighborhood of radius $\lambda_n\delta/2$ about $\underline{S}_n(\underline{\beta}^0)$. Since $\lambda_n\delta/2 \rightarrow \lambda\delta/2$ in probability, there is also an open neighborhood of radius $\lambda\delta/2$ around $\underline{S}_n(\underline{\beta}^0)$, with probability going to one, in the image set $\underline{S}_n(D_\delta)$.

Further, by (A) from Lemma 2, with probability going to one as $n \rightarrow \infty$, we know that $\|\underline{S}_n(\underline{\beta}^0) - 0\| < \lambda\delta/2$, and thus $\underline{0} \in \underline{S}_n(D_\delta)$, with probability going to one as $n \rightarrow \infty$.

Let us now consider the inverse function $\underline{S}_n^{-1} : \underline{S}_n(D_\delta) \rightarrow D_\delta$, which is well defined whenever $\underline{S}_n$ is one-to-one and onto with probability going to one. Finally, since $\underline{0} \in \underline{S}_n(D_\delta)$, with probability going to one as $n \rightarrow \infty$, we may conclude:

1. that the root, $\underline{S}_n^{-1}(\underline{0})$, of Equation 4.7 exists in D_δ , with probability going to one as $n \rightarrow \infty$; this is due to the fact that $\underline{S}_n^{-1}$ exists and is well defined on $\underline{S}_n(D_\delta)$ and therefore may be applied to $\underline{0}$, which is shown to exist in $\underline{S}_n(D_\delta)$, the image set.
2. $\underline{S}_n^{-1}(\underline{0})$ converges in probability to $\underline{\beta}^0$, since δ may be taken to be arbitrarily small; this is because it was just shown in Part 1 that, with probability going to one as $n \rightarrow \infty$, $\underline{S}_n^{-1}(\underline{0})$ exists in D_δ and since δ can be arbitrarily small, $\underline{S}_n^{-1}(\underline{0})$ can get arbitrarily close to the center of D_δ which is $\underline{\beta}^0$.
3. by the one-to-oneness of $\underline{S}_n$ on D_δ , any other sequence $\{\underline{\beta}_n\}$ of solutions to Equation 4.7 necessarily must lie outside of D_δ with probability going to one, i.e., $\{\underline{\beta}_n\}$ does not converge to $\underline{\beta}^0$. This is easily seen since if $\underline{\beta}_n \in D_\delta$ is a solution to Equation 4.7, then $\underline{S}_n(\underline{\beta}_n) = \underline{0}$ and this implies that $\underline{\beta}_n = \underline{S}_n^{-1}(\underline{0})$, as it has been established that $\underline{S}_n^{-1}$ exists on $\underline{S}_n(D_\delta)$. However, if $\hat{\beta}_n \in D_\delta$ is also a solution to Equation 4.7, then we have $\underline{S}_n(\hat{\beta}_n) = \underline{0}$ which implies $\hat{\beta}_n = \underline{S}_n^{-1}(\underline{0})$. The fact that $\underline{S}_n$ is a one-to-one function on D_δ means that the preimages of zero must be the same, i.e., $\hat{\beta}_n = \underline{\beta}_n$, with probability going to one.

The proof is complete upon taking $\hat{\beta}_n = \underline{S}_n^{-1}(\underline{0})$. $\diamond$

The uniqueness established in Theorem 3 is of notable importance not only since any other weakly consistent estimator must be the same (with probability tending to one), but must also, therefore, have the same asymptotic properties, e.g., distribution, as the existent estimator.

The following theorem shows that given a consistent initial estimate, the estimate derived from the modified Gauss-Newton method used in the EGNLS procedure will also be consistent. In addition, the theorem provides a condition which produces a consistent estimate, even when no consistent initial estimate exists.

THEOREM 4: Let a_n be a sequence of constants such that $a_n \rightarrow 0$ as $n \rightarrow \infty$. Suppose that $(\hat{\underline{\beta}}_T - \underline{\beta}^0) = O_p(a_n)$. Let $\hat{\underline{\beta}}$ denote the EGNLS estimator obtained from the modified Gauss-Newton algorithm. Then, under the regularity conditions stated for Lemma 2,

$$(\hat{\underline{\beta}} - \underline{\beta}^0) = O_p(\max(a_n, n^{-1/2})).$$

Further, if it is known that no local minima exist to the quadratic form of Equation 4.5, then the resulting EGNLS estimator from the modified Gauss-Newton algorithm is such that $(\hat{\underline{\beta}} - \underline{\beta}^0) = O_p(n^{-1/2})$, regardless of starting with a consistent estimator or not.

Proof: The proof for the first result of the theorem follows directly that of the analog for a similar theorem in a Ph.D. dissertation by Shah [52, p. 31] for binary variables in diagnostic tests. First, consider the one step estimator $\hat{\underline{\beta}}_k$ of the Gauss-Newton procedure defined in Equation 4.8, having $\hat{\underline{\beta}}_T$ as an initial estimator and $\underline{\beta}^0$ denoting the true parameter value.

Note that $\hat{\underline{\pi}} - \underline{\pi}(\underline{\beta}^0) = O_p(n^{-1/2})$ and $\underline{\pi}(\hat{\underline{\beta}}_T) = \underline{\pi}(\underline{\beta}^0) + O_p(a_n)$ since $(\hat{\underline{\beta}}_T - \underline{\beta}^0) = O_p(a_n)$ and by the continuity of the vector $\underline{\pi}$. Therefore, $[\hat{\underline{\pi}} - \underline{\pi}(\hat{\underline{\beta}}_T)]$ is $O_p(\max(a_n, n^{-1/2}))$. Also, we know $\hat{V}$ is consistent for V and the inverse, V^{-1} , is a continuous function of the elements of V . Further, the matrix $\frac{\partial \underline{\pi}(\underline{\beta})}{\partial \underline{\beta}'}$ is continuous in $\underline{\beta}$ (by regularity condition 2 of Lemma 2), so the term

$$\left[\left(\frac{\partial \underline{\pi}(\hat{\underline{\beta}}_T)}{\partial \underline{\beta}'} \right)' \hat{V}^{-1} \left(\frac{\partial \underline{\pi}(\hat{\underline{\beta}}_T)}{\partial \underline{\beta}'} \right) \right]^{-1} \left(\frac{\partial \underline{\pi}(\hat{\underline{\beta}}_T)}{\partial \underline{\beta}'} \right)' \hat{V}^{-1}$$

converges to

$$\left[\left(\frac{\partial \underline{\pi}(\underline{\beta}^0)}{\partial \underline{\beta}'} \right)' V^{-1} \left(\frac{\partial \underline{\pi}(\underline{\beta}^0)}{\partial \underline{\beta}'} \right) \right]^{-1} \left(\frac{\partial \underline{\pi}(\underline{\beta}^0)}{\partial \underline{\beta}'} \right)' V^{-1},$$

in probability, which is bounded and does not depend on n . Thus, from Equation 4.8 we may write

$$\widehat{\underline{\beta}}_M = \widehat{\underline{\beta}}_T + O_p(\max(a_n, n^{-1/2})),$$

and further,

$$\begin{aligned} (\widehat{\underline{\beta}}_M - \underline{\beta}^0) &= (\widehat{\underline{\beta}}_T - \underline{\beta}^0) + O_p(\max(a_n, n^{-1/2})) \\ &= O_p(\max(a_n, n^{-1/2})). \end{aligned}$$

Now, since each newly iterated estimate also satisfies the same property by similar arguments, the first part of the proof is complete.

The second result of the theorem can trivially be seen by understanding that assuming only a global minimum (to Equation 4.5, i.e., the quadratic form, $Q(\underline{\beta}, \widehat{V})$) exists, then no matter where the modified Gauss-Newton procedure begins, the process must converge to that global minimum. This is due to the fact that the modified Gauss-Newton procedure guarantees convergence of the algorithm. $\diamond$

The key result from Theorem 4 is that the estimator derived using the EGNLS method to estimate the parameters of the GSME model provides consistent estimates assuming a consistent initial estimate exists. A consistent initial estimate for the EGNLS procedure, i.e., for use in the one-step estimation Equation 4.8, is guaranteed, except on a set of probability zero, since Theorem 3 states the existence of a consistent estimator, with probability tending to one as $n \rightarrow \infty$, which is a solution to Equation 4.7. A result of greater importance is that since the EGNLS estimator is consistent, the uniqueness proven in Theorem 3 implies the consistent solution to the EGNLS procedure must have the same asymptotic properties as the weakly consistent estimate that solves Equation 4.7.

The second result of Theorem 4 states a consistent estimate from the modified Gauss-Newton procedure can be reached even if one does not start with a consistent initial estimate, as long as it is known that no local minima to Equation 4.5, i.e., $Q(\underline{\beta}, \widehat{V})$, exist. As stated earlier in the chapter, it is known that the *modified* Gauss-Newton procedure will converge [31, p. 269], which is why it is preferred here over the standard Gauss-Newton procedure, where there may be no guarantee of convergence (see, for example, [31, 64]). Therefore, if only a global minimum exists, the modified Gauss-Newton procedure is guaranteed to find it.

This part of the theorem provides yet another strength to the methods developed for the GSME model. As often is the case in real world situations, one may never be certain if some chosen starting values are consistent for the values of the unknown population parameters. Therefore, in practice, if the existence of a consistent initial estimate is in question, the EGNLS procedure can be performed with multiple starting values from a wide grid, and if those different, not necessarily consistent, starting values result in the same estimate, it can rather safely be assumed that there are no local minima and that all of the initial estimates used for the EGNLS procedure converged to a single global minimum. Note that if local minima exist, then the same conclusion cannot be drawn. The existence of local minima would contradict the uniqueness result from Theorem 3, therefore the resulting estimators would differ (in probability) and thus they would follow different asymptotic distributions and not be applicable to the results given here.

In the next and final section in this chapter, we discuss how consistent initial estimates may be obtained under the GSME model. Since these initial estimates are based on sample proportions, $\widehat{\pi}$, where $\widehat{\pi} - \pi(\beta)$ is $O_p(n^{-1/2})$, using these estimates as starting values, the modified Gauss-Newton procedure will produce an estimator, $\widehat{\underline{\beta}}$, such that $(\widehat{\underline{\beta}} - \underline{\beta}^0) = O_p(n^{-1/2})$.

The following theorem states the large sample distributional properties of solutions to the estimating equations of Equation 4.7 and, thus, are also the properties of the EGNLS estimator. Again, the theorem and proof are based on those by Shah [52, p. 43] for diagnostic tests, but adapted here for the GSME model.

THEOREM 5: Assume that the regularity conditions of Lemma 2 hold and let $\{\hat{\underline{\beta}}_n\}$ be a weakly consistent sequence of estimators that, with probability tending to one as $n \rightarrow \infty$, satisfy Equation 4.7 and $\hat{\underline{\beta}}_n - \underline{\beta}^0 = O_p(n^{-1/2})$. Then

$$\sqrt{n} \left(\hat{\underline{\beta}}_n - \underline{\beta}^0 \right) \xrightarrow{L_{\underline{\beta}^0}} MVN_{\nu}(\underline{0}, U(\underline{\beta}^0)^{-1}),$$

where recall that ν is the dimension of $\underline{\beta}^0$.

Proof: Given $\varepsilon > 0$, Theorem 3 implies that, with probability going to one as $n \rightarrow \infty$, there exists an n_0 such that for $n > n_0$, $\{\hat{\underline{\beta}}_n\} \in D_\delta$ and $\{\hat{\underline{\beta}}_n\}$ satisfies Equation 4.7 with $\lim_{n \rightarrow \infty} P\left(\underline{S}_n(\hat{\underline{\beta}}_n) = \underline{0}\right) = 1$. Now, using a Taylor Series (see, for example, [50, p. 17]) expand the vector $\underline{S}_n(\hat{\underline{\beta}}_n)$ about $\underline{\beta}^0$, and thus for $\{\hat{\underline{\beta}}_n\} \in D_\delta$, we have

$$\begin{aligned} \underline{0} &= \underline{S}_n(\hat{\underline{\beta}}_n) \\ &= \underline{S}_n(\underline{\beta}^0) + S'_n(\underline{\beta}^0) \left(\hat{\underline{\beta}}_n - \underline{\beta}^0 \right) + O_p(n^{-1}), \end{aligned}$$

with probability tending to one as $n \rightarrow \infty$. Therefore,

$$P_{\underline{\beta}^0} \left[\lim_{n \rightarrow \infty} \left| n^{1/2} \underline{S}_n(\underline{\beta}^0) + n^{1/2} S'_n(\underline{\beta}^0) \left(\hat{\underline{\beta}}_n - \underline{\beta}^0 \right) \right| = 0 \right] = 1.$$

By results (B) and (C) of Lemma 2

$$P_{\underline{\beta}^0} S'_n(\underline{\beta}^0) \xrightarrow{L_{\underline{\beta}^0}} -U(\underline{\beta}^0),$$

which implies

$$P_{\underline{\beta}^0} \left[\lim_{n \rightarrow \infty} \left| n^{1/2} \underline{S}_n(\underline{\beta}^0) + n^{1/2} [-U(\underline{\beta}^0)] \left(\widehat{\underline{\beta}}_n - \underline{\beta}^0 \right) \right| = 0 \right] = 1,$$

and, thus,

$$P_{\underline{\beta}^0} \left[\lim_{n \rightarrow \infty} \left| n^{1/2} [-U(\underline{\beta}^0)^{-1}] \underline{S}_n(\underline{\beta}^0) + n^{1/2} \left(\widehat{\underline{\beta}}_n - \underline{\beta}^0 \right) \right| = 0 \right] = 1.$$

Therefore, we know that, with probability going to one as $n \rightarrow \infty$, the limiting distribution of $n^{1/2} \left(\widehat{\underline{\beta}}_n - \underline{\beta}^0 \right)$ is the same as that of $n^{1/2} [-U(\underline{\beta}^0)^{-1}] \underline{S}_n(\underline{\beta}^0)$. Using the Multivariate Central Limit Theorem [45, p. 128] and the Delta Method (see, for example, [1, p. 56-58]) yields

$$n^{1/2} \underline{S}_n(\underline{\beta}^0) \xrightarrow{L_{\underline{\beta}^0}} MVN_P(\underline{0}, U(\underline{\beta}^0)).$$

The result, then, follows. $\diamond$

Note that after the last step in the modified Gauss-Newton algorithm, the asymptotic variance-covariance matrix of the EGNLS estimates can be directly estimated. Let $\widehat{\underline{\beta}}$ represent the EGNLS estimators, then the approximate asymptotic variance-covariance matrix of $\widehat{\underline{\beta}}$ after the last step of the Gauss-Newton algorithm is estimated by

$$\left[\left(\frac{\partial \pi(\widehat{\underline{\beta}})}{\partial \underline{\beta}'} \right)' \widehat{V}^{-1} \left(\frac{\partial \pi(\widehat{\underline{\beta}})}{\partial \underline{\beta}'} \right) \right]^{-1}.$$

It was shown by Shah [52, p. 44] that for diagnostic tests with binary variables, that the EGNLS estimators are asymptotically efficient when compared to the maximum likelihood estimators. We cannot arrive at that same result here. In applying EGNLS to the GSME model, we know that efficiency has been lost due to the categorization of manifest variables and categorization of x . We shall discuss possible future work concerning the asymptotic efficiency of the estimators for the GSME model in the final chapter of this dissertation.

4.4 Consistent Initial Estimates

In the previous section, theorems were stated and proven which gave the results that the EGNLS method produced parameter estimators which were consistent and asymptotically normal for the GSME model, given that consistent initial estimates are used in the estimation procedure. Recall that the vector of model parameters is

$$\underline{\beta} = (\underline{\beta}'_1, \underline{\beta}'_2, \dots, \underline{\beta}'_p, \underline{\theta}'_x)'$$

In this section we describe a method to derive consistent estimates to be used as initial trial values, $\widehat{\underline{\beta}}_T$, for use in EGNLS.

By taking a just-identified subset of invertible equations from Equation 4.4, i.e., $\widehat{\underline{\pi}} = \underline{\pi}(\underline{\beta}) + \underline{\varepsilon}$, one can solve for the elements of $\underline{\beta}$ in terms of sample proportions, $\widehat{\underline{\pi}}$, which are consistent estimates of the true cell probabilities, $\underline{\pi}$. The dimension of $\underline{\pi}$ is $N - 1$, so as long as the number of elements in $\underline{\beta}$, which is equal to $\nu = (l_1 + l_2 + \dots + l_p + k_x)$, is less than N we can take a just-identified subset of equations and solve for our initial estimates which will be consistent. Recall that the calculation of N may vary for different applications.

This method of using a just-identified subset of equations to derive initial estimates relates directly to the use of an invertible subset of equations used in Section 3.3 of Chapter 3 on identifiability. Theorem 2 of that section establishes the identifiability of the GSME model through the existence of an invertible subset

of equations from the parameters $\underline{\omega}(\underline{\beta})$ of the joint distribution of the manifest variables, $f_Y(y; \underline{\omega}(\underline{\beta}))$. Recall that Theorem 2 stated that $\underline{\beta}$ is identified, and thus the GSME model is identified, if there exists a proper subvector of $\underline{\omega}$, say $\underline{\omega}_1$, such that $\underline{\omega}_1^{-1}(\underline{\omega}_1) = \underline{\beta}$. This means that when such an $\underline{\omega}_1$ exists, we know an invertible, proper subset of $\underline{\pi}$ does as well. This is because $\underline{\pi}$ is the vector of parameters of the joint marginal distribution of the manifest variables (given by $\underline{\omega}$ in Section 3.3 on identification of the GSME model), after categorization.

CHAPTER 5

APPLICATION TO MOTIVATING EXAMPLE

In this chapter we show how the motivating example from the RERF in Hiroshima, Japan, which was first discussed in Chapter 1, can be modified to fit into the GSME model framework discussed in Chapter 3 and how the parameters of the associated models can be estimated using the procedure described in Chapter 4. Let Y_1 , the outcome of interest, be the time until the diagnosis of leukemia following A-bomb radiation exposure. The variable contaminated by measurement error is the amount of true radiation dosage an individual was exposed to, it is denoted by x , and it is estimated from the DS86 system. DS86 stands for the Dosimetry System 1986 and it estimates an individuals true, unknown level of radiation exposure, where the imprecision in an individuals estimate comes primarily through uncertainties about survivors' shielding and location to the specific bombs' hypocenter [42, p. 275]. The DS86 estimate is denoted by Y_2 . The identifying information comes through the instrumental variable, which is an indicator of epilation, and this shall be denoted as Y_3 . Epilation is severe hair loss due to radiation exposure (scalp hair loss of more than 67% [41]) and here a value of 1 indicates epilation and a value of 0 indicates no epilation. We shall heuristically assume that the parameters for this GSME model are identified. Further, for the current application, we shall assume that there is no censored data. Showing how the outcome, Y_1 , can be represented by a GSME model is not trivial, so we shall save that task for later in the chapter and start with the remaining variables, first.

Our assumptions follow closely those used by Pierce et al. [43] which have been established by the RERF from previous research [42]. The distributions

involved in our assumptions in this chapter are discussed in more detail by Kalbfleisch and Prentice [33]. We assume that x follows a Weibull distribution where its units are measured in grays of radiation. This is denoted as $x \sim \text{Weibull}(\underline{\theta}_x)$, where $\underline{\theta}_x = (a, b)$ and $f(x \mid \underline{\theta}_x) = abx^{b-1} \exp(-ax^b)$. Estimated Radiation dose, Y_2 , is assumed to follow a log-normal distribution. Therefore, we have $\log(Y_2) \sim N(\mu_2, \sigma_2^2)$, where $\mu_2 = \log(x) = h_{21}(x; \underline{\beta}_{21})$, in the GSME model notation from Equation 3.4. That is, $\log(Y_2) = \log(x) + u$, where $u \sim N(0, \sigma_2^2)$. Then, conditional on x , the log-normal density for Y_2 is written as $f_2(Y_2 \mid x, \underline{\theta}_2) = (2\pi)^{-1/2} p_2 y_2^{-1} \exp(-[\log(\lambda_{21} y_2)]^2 p_2^2 / 2)$, where $\underline{\theta}_2 = (\lambda_{21}, p_2)$, $\lambda_{21} = 1/x$ and $p_2 = 1/\sigma_2$. The assumptions for the instrumental variable are as follows: $Y_3 \sim \text{Bernoulli}(\theta_3)$, where $\mu_3 = h_{31}(x; \underline{\beta}_{31}) = \exp(\beta_{30} + \beta_{31}x) / [1 + \exp(\beta_{30} + \beta_{31}x)]$ and the conditional distribution written in terms of the conditional distribution parameters is $f_3(Y_3 \mid x; \underline{\theta}_3) = \theta_3^{y_3} (1 - \theta_3)^{1-y_3}$, where $\underline{\theta}_3 = \theta_3$.

Now that we have seen how the other manifest variables and x can be defined by the mappings of the GSME model, let us consider the outcome of interest; time until the diagnosis of leukemia. Let $\lambda(y_1, x)$ denote the hazard function of Y_1 , given radiation dose x . That is, given dose x ,

$$\lambda(y_1, x) = \lim_{\Delta y_1 \rightarrow 0} \text{Pr}(y_1 \leq Y_1 < y_1 + \Delta y_1 \mid Y_1 \geq y_1, x) / \Delta y_1$$

and is interpreted as the rate of instantaneous failure (e.g., diagnosis of leukemia) after time y_1 given that an individual who received dose x is disease free at y_1 . It is well known (see [33, p. 6]) that

$$f_1(y_1 \mid x, \underline{\theta}_1) = \lambda(y_1, x) \exp \left(- \int_0^{y_1} \lambda(u, x) du \right).$$

Therefore, one can specify $\lambda(y_1, x)$ to define $f_1(y_1 \mid x, \underline{\theta}_1)$. Note that usually the independent variable in a hazard function is time and therefore denoted by T ,

however, we retain Y_1 to avoid confusion with the T mappings from the equations in 3.2 in Chapter 3.

Individuals who were in Hiroshima or Nagasaki during the times that the bombs were dropped could have developed leukemia which was not induced by the exposure to this radiation. It is assumed here that the chances of developing a non-radiation induced leukemia exists at the beginning of the study and is constant throughout, never increasing or decreasing. This can be thought of as a steady, baseline or background hazard rate, which is always present in every individual. This baseline hazard rate is denoted by λ_0 and does not depend on y_1 or x , and is therefore constant. The fact that a non-radiation induced leukemia could occur for any individual, even though they were still exposed to radiation from the bombs, must be accounted for in the hazard plot. Therefore, the hazard plot for Y_1 cannot start with an initial hazard rate for an individual of zero because there is some initial chance that every individual in the study could develop a non-radiation induced leukemia. Assume a plot of $\lambda(y_1, x)$ appears as illustrated in Figure 5-1.

We could, as an example, specify $\lambda(y_1, x)$ as,

$$\lambda(y_1, x) = \lambda_0 R_1(y_1, x),$$

where λ_0 , again, is the baseline hazard rate and $R_1(y_1, x)$ is the relative risk (henceforth, RR) function satisfying $R_1(0, x) = R_1(y_1, 0) = 1$, and $R_1(\infty, x) = 1$. For example, $R_1(y_1, x)$ might be $R_1(y_1, x) = \exp[x \exp(\beta_{11}y_1 + \beta_{12} \log y_1)]$, $\beta_{11} < 0$ and $\beta_{12} > 0$. Or, we might specify $\lambda(y_1, x)$ to be

$$\lambda(y_1, x) = \lambda_0 + R_1^*(y_1, x),$$

where $R_1^*(y_1, x)$ is the attributable risk (henceforth, AR) function satisfying $R_1^*(0, x) = R_1^*(y_1, 0) = 0$ and $R_1^*(\infty, x) = 0$. As an example, $R_1^*(y_1, x)$ might be

$$R_1^*(y_1, x) = x \exp(\beta_{11}y_1 + \beta_{12} \log y_1) = xy_1^{\beta_{12}} \exp(\beta_{11}y_1),$$

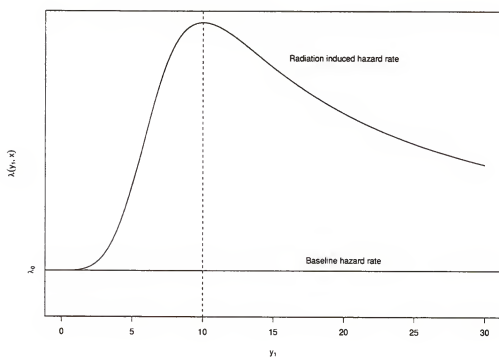


Figure 5-1: Hazard Plot for Time Until Diagnosis of Leukemia After A-Bomb Radiation Exposure

where $\beta_{11} < 0$ and $\beta_{12} > 0$.

Modeling Y_1 by using either of these two forms for its hazard function, $\lambda(y_1, x)$, would not fit into the GSME model framework because these parameters are functions of both y_1 and x . In the RERF study, the hazard function for $f_1(y_1 | x, \underline{\theta}_1)$ cannot be modeled by a parameter which uses a RR or an AR function because to fit into the GSME model framework the parameter must be a function of x only; the unobserved level of radiation exposure. This parameter is seen in the mappings which define the GSME model as $h_{11}(x; \underline{\beta}_{11})$, a function of x (and $\underline{\beta}_{11}$) only. So, at first glance, it might appear that this problem is outside of the scope of the GSME model. It turns out that some survival analysis problems may be modified in such a way as to include a parameter which is a function of x alone, and therefore, in the modified form, meet the requirements that a GSME model must satisfy. In this specific application, writing the hazard function for Y_1 as the sum of two separate parts offers a solution to the problem. The hazard function in this case can be separated into two independent components: one component representing the instantaneous rate of failure (diagnosis of leukemia) based solely on radiation exposure and the other component specifying the instantaneous rate of failure based on the baseline, or background, rate for developing a non-radiation induced leukemia. In what follows, we show how this example can be modified to fit into the Latent Class model developed in Chapter 4, due to the fact that one of these two separate components can be modeled to include a parameter which is a function of x alone.

To show how the RERF problem fits into the GSME model framework, recall that Y_1 is the time to diagnosis of leukemia after exposure to radiation. This may actually be modeled by combining two separate hazards for the two types of leukemia which may be diagnosed: a radiation induced leukemia or a non-radiation induced leukemia. Let Y_{11} be the time until the diagnosis of a leukemia that was

due to radiation exposure and Y_{12} be the time until the diagnosis of a leukemia in a case that was not induced by radiation exposure. Then $Y_1 = \min(Y_{11}, Y_{12})$. Assume that Y_{11} and Y_{12} are independent. Further, assume that Y_{12} has a hazard function denoted by $\lambda_2(y_1)$, which is free of x , and Y_{11} has a hazard function denoted by $\lambda_1(y_1, x)$ with a shape illustrated by the top curve only, labeled "Radiation induced hazard rate", in Figure 5-1, with $\lambda_1(y_1, x) = \lambda_1(y_1, \lambda(x))$, where $\lambda_1(y_1, \lambda(0)) = 0$. Then the density function of Y_{11} given x has a parameter that is a function of x , namely $\lambda(x)$, since $f_1(y_{11} | x, \underline{\theta}_1) = \lambda_1(y_{11}, \lambda(x)) \exp(-\int_0^{y_{11}} \lambda_1(u, \lambda(x)) du)$.

Now the hazard function of Y_1 given x is

$$\begin{aligned}
 \lambda(y_1, x) &= \lim_{\Delta y_1 \rightarrow 0^+} P(y_1 \leq Y_1 < y_1 + \Delta y_1 | Y_1 \geq y_1) / \Delta y_1 \\
 &= \lim_{\Delta y_1 \rightarrow 0^+} [P(y_1 \leq Y_1 < y_1 + \Delta y_1 \cap Y_{11} \geq y_1 \cap Y_{12} \geq y_1) / P(Y_{11} \geq y_1, Y_{12} \geq y_1)] / \Delta y_1 \\
 &= \lim_{\Delta y_1 \rightarrow 0^+} [P\{(Y_{11} < y_1 + \Delta y_1 \cup Y_{12} < y_1 + \Delta y_1) \cap (Y_{11} \geq y_1, Y_{12} \geq y_1)\} / P(Y_1 \geq y_1) \Delta y_1] \\
 &= \lim_{\Delta y_1 \rightarrow 0^+} [P\{(y_1 \leq Y_{11} < y_1 + \Delta y_1 \cap Y_{12} \geq y_1) \cup (y_1 \leq Y_{12} < y_1 + \Delta y_1 \cap Y_{11} \geq y_1)\} / P(Y_1 \geq y_1) \Delta y_1] \\
 &= \lim_{\Delta y_1 \rightarrow 0^+} [P\{y_1 \leq Y_{11} < y_1 + \Delta y_1 \cap Y_{12} \geq y_1\} + P\{y_1 \leq Y_{12} < y_1 + \Delta y_1 \cap Y_{11} \geq y_1\} - \\
 &\quad P\{y_1 \leq Y_{11} < y_1 + \Delta y_1 \cap Y_{12} \geq y_1 \cap y_1 \leq Y_{12} < y_1 + \Delta y_1 \cap Y_{11} \geq y_1\} / P(Y_{11} \geq y_1) P(Y_{12} \geq y_1) \Delta y_1] \\
 &= \lambda_1(y_1, \lambda(x)) + \lambda_2(y_1) - \lim_{\Delta y_1 \rightarrow 0^+} \left[\frac{P(y_1 \leq Y_{11} < y_1 + \Delta y_1 | Y_{11} \geq y_1)}{\Delta y_1} \right] \\
 &\quad \times \left[\frac{P(y_1 \leq Y_{12} < y_1 + \Delta y_1 | Y_{12} \geq y_1)}{\Delta y_1} \right] \Delta y_1 \\
 &= \lambda_1(y_1, \lambda(x)) + \lambda_2(y_1).
 \end{aligned}$$

This implies that

$$f_1(y_1 | x, \underline{\theta}_1) = f(y_{11} | x, \underline{\theta}_1) S(y_{12} | x, \underline{\theta}_1) + f(y_{12} | x, \underline{\theta}_1) S(y_{11} | x, \underline{\theta}_1),$$

where $S(\cdot)$ is the survival function, e.g.,

$$S(y_{12} | x, \underline{\theta}_1) = \int_{y_{12}}^{\infty} f(u) du = \exp\left(-\int_0^{y_{12}} \lambda_2(u) du\right),$$

and this can easily be shown using the fact that

$$f_1(y_1 | x, \underline{\theta}_1) = \lambda(y_1, x) \exp\left(-\int_0^{y_1} \lambda(u, x) du\right).$$

This density is a function of the parameters of $f(y_{11} | x, \underline{\theta}_1)$ and $f(y_{12} | x, \underline{\theta}_1)$. By our assumptions, this includes a parameter $\lambda(x)$.

Now $f_1(y_1 | x, \underline{\theta}_1)$ can be specified by choosing any density for Y_{12} and any density for Y_{11} with a hazard function of the previously illustrated shape, by setting one parameter of $f(y_{11} | x, \underline{\theta}_{11})$ to $\lambda(x)$ and specifying a functional form for $\lambda(x)$.

In light of this new way of expressing the density for Y_1 to fit into the GSME model framework, let us return to the RERF example specifications. Assume that $Y_{11} \sim \text{Log} - \text{logistic}(\lambda_{11}, p_1)$ and that $\lambda_{11} = \beta x$, where $\beta > 0$. Then, we have, $f(y_{11} | x, \underline{\theta}_{11}) = \lambda_{11} p_1 (\lambda_{11} y_1)^{p_1-1} [1 + (\lambda_{11} y_1)^{p_1}]^{-2}$ and $\lambda_1(y_1, \lambda(x)) = (\beta x) p_1 ((\beta x) y_1)^{p_1-1} / [1 + ((\beta x) y_1)^{p_1}]$, a hazard function having the desired shape. Further assume that $Y_{12} \sim \exp(\lambda_{12})$, i.e., is exponentially distributed. Then $f(y_{12} | x, \underline{\theta}_{12}) = \lambda_{12} \exp(-\lambda_{12} y_1)$ and $\lambda_2(y_1) = \lambda_{12}$. So, combining these, we have

$$f_1(y_1 | x, \underline{\theta}_1) = \lambda_{12} \exp(-\lambda_{12} y_1) [1 + (\lambda_{11} y_1)^{p_1}]^{-1} + \lambda_{11} p_1 (\lambda_{11} y_1)^{p_1-1} [1 + (\lambda_{11} y_1)^{p_1}]^{-2} \exp(-\lambda_{12} y_1)$$

as the density for Y_1 in the GSME model.

So, for this specification, we have that the distributional parameters are $\underline{\theta}_1 = (\lambda_{11}, p_1, \lambda_{12})'$, $\underline{\theta}_2 = (\lambda_{21}, p_2)'$, and $\theta_3 = \mu_3$ and therefore we know that $k_2 = 2$, $k_1 = 3$, and $k_3 = 1$. For the functions of the distributional parameters which are used with Equation 3.4, we have that T_1 is the identity mapping, the mapping T_2 consists of two elements and they are defined by $T_{21}(\lambda_{21}, p_2) = -\log(\lambda_{21}) = \mu_2$ and $T_{22}(\lambda_{21}, p_2) = 1/p_2^2 = \sigma_2^2$ (so $\underline{\gamma}_2 = (\mu_2, \sigma_2^2)'$), and the mapping T_3 is also the identity, i.e., $T_3(\theta_3) = \mu_3 = \theta_3$. To complete the specification of parameters found in Equation 3.4, let $h_{11}(x; \beta_{11}) = \lambda_{11} = \beta x$, $h_{12}(x; \beta_{12}) = p_1$, and $h_{13}(x; \beta_{13}) = \lambda_{12}$, for Y_1 , $h_{21}(x; \beta_{21}) = \mu_2 = \log(x)$ and $h_{22}(x; \beta_{22}) = \sigma_2^2$, for Y_2 , and $h_{31}(x; \beta_{31}) = \theta_3 = \mu_3 = \exp(\beta_{30} + \beta_{31} x) / [1 + \exp(\beta_{30} + \beta_{31} x)]$, for Y_3 .

Now we have the needed densities to allow us to write Equation 4.1 for this application. Recall that Equation 4.1 is the joint distribution of the manifest

variables, prior to the categorization step. We have

$$\begin{aligned}
 f(y_1, y_2, y_3) &= \int_x \lambda_{12} \exp(-\lambda_{12} y_1) [1 + (\beta x y_1)^{p_1}]^{-1} \\
 &\quad + \beta x p_1 (\beta x y_1)^{p_1-1} [1 + (\beta x y_1)^{p_1}]^{-2} \exp(-\lambda_{12} y_1) \\
 &\quad \times (2\pi)^{-1/2} (\sigma_2 y_2)^{-1} \exp(-[\log(y_2) - \log(x)]^2 / (2\sigma_2^2)) \\
 &\quad \times \left[\frac{\exp(\beta_{30} + \beta_{31} x)}{1 + \exp(\beta_{30} + \beta_{31} x)} \right]^{y_3} \left[\frac{1}{1 + \exp(\beta_{30} + \beta_{31} x)} \right]^{1-y_3} \\
 &\quad \times a b x^{b-1} \exp(-a x^b) dx.
 \end{aligned}$$

We categorize Y_1 , Y_2 , and x and the analog, then, to the left hand side of the previous equation is $\pi_{j_1 j_2 j_3}$, a multinomial cell probability, and the analogue to the right hand side is

$$\begin{aligned}
 \sum_{j_x} P(Y_{1i} \in I_{1j_1} \mid x_i \in I_{j_x}, \underline{\theta}_1) P(Y_{2i} \in I_{2j_2} \mid x_i \in I_{j_x}, \underline{\theta}_2) \times \\
 P(Y_{3i} \in I_{3j_3} \mid x_i \in I_{j_x}, \underline{\theta}_3) P(x_i \in I_{j_x} \mid \underline{\theta}_x),
 \end{aligned}$$

where

$$\begin{aligned}
 P(Y_{1i} \in I_{1j_1} \mid x_i \in I_{j_x}, \underline{\theta}_1) &= S(l_{1j_1 i}) - S(u_{1j_1 i}) \\
 P(Y_{2i} \in I_{2j_2} \mid x_i \in I_{j_x}, \underline{\theta}_2) &= S(l_{2j_2 i}) - S(u_{2j_2 i}) \\
 P(Y_{3i} = 1 \mid x_i \in I_{j_x}, \underline{\theta}_3) &= \frac{\exp(\beta_{30} + \beta_{31} x_{j_x})}{1 + \exp(\beta_{30} + \beta_{31} x_{j_x})} \\
 P(Y_{3i} = 0 \mid x_i \in I_{j_x}, \underline{\theta}_3) &= \frac{1}{1 + \exp(\beta_{30} + \beta_{31} x_{j_x})} \\
 P(x \in I_{j_x} \mid \underline{\theta}_x) &\doteq a b x_{j_x}^{b-1} \exp(-a x_{j_x}^b) \times [u_{j_x} - l_{j_x}].
 \end{aligned}$$

One adjustment that could be made to our assumptions is that Y_2 given x could follow a log-logistic distribution, instead of a log-normal distribution. In this case we would have $Y_2 \sim \text{Log-logistic}(\lambda_{21}, p_2)$. The advantage to this assumption is that the CDF of the log-logistic distribution exists in a closed form, and therefore its survival function does as well, whereas that of the log-normal does not. There are no real disadvantages to making this change in assumptions,

because under the log-logistic assumption, $\log(Y_2)$ follows a logistic distribution which is symmetric with only slightly heavier tails than a normal distribution. Therefore, in practice, the log-normal and log-logistic distributional assumptions for Y_2 given x are essentially interchangeable.

In the case where it is assumed that $Y_2 \sim \text{Log-logistic}(\lambda_{21}, p_2)$, where $\lambda_{21} = 1/x$ and $E(\log Y_2) = \log x$. Then, for the survival function, we would have $S(y_2) = 1/(1 + (\lambda_{21}y_2)^{p_2})$. Furthermore, for Y_1 we have that $S(y_1) = S(y_{11})S(y_{12})$, where $S(y_{11}) = 1/(1 + (\lambda_{11}y_{11})^{p_1})$ and $S(y_{12}) = \exp(-\lambda_{12}y_{12})$. So, after categorization, for the latent class model Equation 4.3, we would have for the $(j_1 = 1, j_2 = 1, j_3 = 1)$ th multinomial cell probability

$$\begin{aligned} \pi_{111} &\doteq \sum_{j_x} \left[\frac{\exp(-\lambda_{12}l_{11})}{1 + (\beta x_{j_x}(l_{11}))^{p_1}} - \frac{\exp(-\lambda_{12}u_{11})}{1 + (\beta x_{j_x}(u_{11}))^{p_1}} \right] \\ &\times \left[\frac{1}{1 + ((1/x_{j_x})l_{21})^{p_2}} - \frac{1}{1 + ((1/x_{j_x})u_{21})^{p_2}} \right] \\ &\times \left[\frac{\exp(\beta_{30} + \beta_{31}x_{j_x})}{1 + \exp(\beta_{30} + \beta_{31}x_{j_x})} \right] \\ &\times abx_{j_x}^{b-1} \exp(-ax_{j_x}^b) \times [u_{j_x} - l_{j_x}] \\ &\doteq \pi_{111}(\underline{\beta}), \end{aligned}$$

where $j_3 = 1$ is equivalent to $Y_{3i} = 1$ and $\underline{\beta} = (\beta, p_1, \lambda_{12}, p_2, \beta_{30}, \beta_{31}, a, b)'$ is the vector of model parameters which entered the conditional densities through the mappings defined in Equation 3.4. In general, the probability of falling into cell $j_1j_2j_3$, where $j_r = 1, 2, \dots, C_r$, $r = 1, 2, 3$, of the multinomial distribution is

$$\begin{aligned} \pi_{j_1j_2j_3}(\underline{\beta}) &\doteq \sum_{j_x} \left[\frac{\exp(-\lambda_{12}l_{1j_1})}{1 + (\beta x_{j_x}(l_{1j_1}))^{p_1}} - \frac{\exp(-\lambda_{12}u_{1j_1})}{1 + (\beta x_{j_x}(u_{1j_1}))^{p_1}} \right] \\ &\times \left[\frac{1}{1 + ((1/x_{j_x})l_{2j_2})^{p_2}} - \frac{1}{1 + ((1/x_{j_x})u_{2j_2})^{p_2}} \right] \\ &\times \left[\frac{1}{1 + \exp(\beta_{30} + \beta_{31}x_{j_x})} \right]^{1-y_{3j_3}} \left[\frac{\exp(\beta_{30} + \beta_{31}x_{j_x})}{1 + \exp(\beta_{30} + \beta_{31}x_{j_x})} \right]^{y_{3j_3}} \\ &\times abx_{j_x}^{b-1} \exp(-ax_{j_x}^b) \times [u_{j_x} - l_{j_x}]. \end{aligned}$$

CHAPTER 6 OTHER POSSIBLE APPLICATIONS

6.1 Application for MCHERDC with Developmental Delay and Disability Outcome

The Maternal, Child Health and Education Research and Data Center (henceforth, MCHERDC), in the College of Medicine at the University of Florida has access to many state-wide data sets to perform epidemiological research studies in the area of maternal and child health. In a future study, we will apply the methods presented in this dissertation to data from MCHERDC. The outcome variable, Y_1 , will be an indicator (Yes/No) of early childhood developmental delay or disability (henceforth, DDD). The mean of this manifest variable will be modeled as a function of the true gestational age (henceforth, GA), which is denoted as x .

The outcome of interest is an indicator of DDD, so per each individual, we know that Y_1 follows a Bernoulli distribution where we are interested in modeling the mean, i.e., the proportion of “successes” or DDD’s. We assume a logistic mean function. Therefore, the mapping for this variable defined in Equation 3.1 that we are interested in modeling is $\gamma_{11} = h_{11}(x; \underline{\beta}_1) = \frac{\exp(\beta_{10} + \beta_{11}x)}{1 + \exp(\beta_{10} + \beta_{11}x)}$. So $\underline{\beta}_1 = (\beta_{01}, \beta_{11})'$, where β_{01} and β_{11} are the intercept and slope, respectively, of the linear predictor in a logistic regression model. The mapping related to γ_{11} is written in terms of the conditional distribution parameters, that is the mapping defined in Equation 3.2, is $\gamma_{11} = P(Y_1 = 1 | x) = h_{11}(x; \underline{\beta}_1)$, where $Y_1 = 1$ if a DDD is observed, zero otherwise. Now, because $\underline{\theta}_1 = E(Y_1 | x)$, where $\underline{\theta}_1 = P(Y_1 = 1 | x)$, the density in terms of the conditional distribution parameter is then $f(y_1 | x, p) = p^{y_1}(1 - p)^{1-y_1}$, when Y_1 is Bernoulli. Since this variable is already dichotomous, there is no need

for further categorization during the estimation procedure for the GSME model. Assuming x has been categorized, the probabilities used in developing the right hand side of the multinomial relation $\hat{\pi} = \pi(\beta) + \varepsilon$, i.e., Equation 4.4, for this variable would simply be

$$\begin{aligned} P(Y_1 = 1 \mid x \in I_{j_x}, \theta_1) &= \frac{\exp(\beta_{01} + \beta_{11}x_{j_x})}{1 + \exp(\beta_{01} + \beta_{11}x_{j_x})} \\ P(Y_1 = 0 \mid x \in I_{j_x}, \theta_1) &= \frac{1}{1 + \exp(\beta_{01} + \beta_{11}x_{j_x})}. \end{aligned}$$

The $\beta_{\perp 1}$ model parameters are not identified, of course, without additional information. We now discuss its estimation using an instrumental variable.

There are many different ways to estimate true GA, but the one assumed to be most reliable is the estimate that is based on the mother's last menstrual period, or LMP [39], i.e., $GA = DOB - (LMP + 14)$. This estimate based on LMP will serve as Y_2 , the measured value of x , true GA, in this application. It is reasonable to assume that $Y_2 = x + u$, where $E(u) = 0$, and that Y_2 is logistically distributed (a symmetric distribution with only slightly heavier tails than the normal distribution).

There are a number of possibilities to consider for Y_3 , the instrumental variable. One very easily accessible variable is birth weight (BW). Whether BW meets the requirements of an IV, however, is debatable. While there is no question that BW is independent of the error in Y_2 , i.e., the measurement error, the independence between BW and the model error in Y_1 is questionable. One school of thought is that given true GA, BW may not have an effect on DDD, i.e., at a fixed value for true GA, there may not be a correlation between BW and DDD. If this assumption were true, BW would satisfy the criteria for an IV and its accessibility in current data sets would make it a prime candidate as the IV in this application. A second variable that is easily accessible and might serve as an IV is length of stay (LOS), which is the amount of time an infant remains in the hospital

after birth. This variable is clearly correlated with true GA, as a very premature infant would most certainly have a long LOS after birth. Further, LOS is not used in estimating GA, thus LOS is clearly uncorrelated with the measurement error in the estimated value of GA. The assumption would have to be made that LOS is independent of the model error in the outcome, DDD. This is the variable considered for the IV at the end of this section, where some details from Chapters 3 and 4 are given in the context of this example.

Other possible choices for Y_3 come from other independent estimates of GA. There are at least three alternative methods for estimating GA. All of these estimates come from neonatal (immediately after birth) methods and therefore would be independent of the measurement error in estimating true GA through LMP. One problem is that these variables are not immediately accessible in the most commonly used data sets by MCHERDC, thus some merging of data sets would be required prior to analysis. The three other methods of estimation for GA are the Ballard score, which is from measurements on the fetus taken immediately after birth (see [39, 53]), an estimate based on the biparietal diameter, i.e., a head measurement [37], or an estimate based on sonographic measurement of femur length [39]. Further research or discussions with practitioners, as well as determining accessibility of these variables, would be required to determine which of these would serve best as the IV, Y_3 .

Once these initial issues are resolved, the first step in estimating the model parameters will be to categorize all continuous variables and develop an analogous equation to Equation 4.3, the extended Latent Class Model analog to the joint distribution of the manifest variables, but in terms of the variables specific to this application. As was seen above, Y_1 is already a categorical manifest variable, and we know that Y_2 , Y_3 , and x would follow continuous distributions that would require the categorization described in Chapter 4. The joint marginal distribution

of Y_1 , Y_2 , and Y_3 , after categorization, is, again, a multinomial distribution. The starting point for the probability of falling into cell $j_1j_2j_3$ of the cross classification of Y_1 and the remaining categorized variables is written as

$$\begin{aligned} & \pi_{j_1j_2j_3}(\underline{\beta}) \doteq \\ & \sum_{j_x} \left[\left[\frac{\exp(\beta_{10} + \beta_{11}x)}{1 + \exp(\beta_{10} + \beta_{11}x)} \right]^{j_1} \left[\frac{1}{1 + \exp(\beta_{10} + \beta_{11}x)} \right]^{1-j_1} \right. \\ & \times \left. \int_{I_{2j_2}} f_2(y_2 | x = x_{j_x}, \underline{\beta}_2) dy_2 \int_{I_{3j_3}} f_3(y_3 | x = x_{j_x}, \underline{\beta}_3) dy_3 f(x_{j_x} | \underline{\theta}_x) L_{j_x} \right], \quad (6.1) \end{aligned}$$

where j_1 is zero or one, for no DDD or observing a DDD, respectively.

Let us now consider the remaining variables. Assume that true GA, x , follows a Weibull distribution, Y_2 , the estimated value of GA, follows a Logistic distribution (as mentioned above), and that for Y_3 , the IV, we consider LOS and assume it follows a Log-Logistic distribution. These distribution assumptions are similar to those made by Resnick et al. [47]. As mentioned above, LOS is clearly related to true GA and is not used in estimating GA, therefore is it independent of the measurement error. We must heuristically assume LOS is uncorrelated with the model error in Y_1 .

There are no mappings required for x in the GSME model, therefore, assuming $x \sim \text{Weib}(a, b)$, we have $P(x \in I_{j_x} | \underline{\theta}_x) \doteq abx_{j_x}^{b-1} \exp(-ax_{j_x}^b) [L_{j_x}]$ for the j_x th category of x . This is the required piece for $f(x_{j_x} | \underline{\theta}_x) L_{j_x}$ in Equation 6.1. Assuming $Y_2 \sim \text{logistic}(\lambda_2, p_2)$, and is an unbiased measurement of GA, the mappings defined in Chapter 3 for this variable would be

$$\begin{aligned} & \begin{bmatrix} \gamma_{21} \\ \gamma_{22} \end{bmatrix} = \begin{bmatrix} h_{21}(x; \underline{\beta}_{21}) \\ h_{22}(x; \underline{\beta}_{22}) \end{bmatrix} = \begin{bmatrix} x \\ \beta_{22} \end{bmatrix} \\ \Leftrightarrow & \begin{bmatrix} \gamma_{21} \\ \gamma_{22} \end{bmatrix} = \begin{bmatrix} T_{21}(\underline{\theta}_2) \\ T_{22}(\underline{\theta}_2) \end{bmatrix} = \begin{bmatrix} \lambda_2 \\ p_2^2 \pi^2 / 3 \end{bmatrix}. \end{aligned}$$

A closed form CDF exists for this distribution so for $\int_{I_{2j_2}} f_2(y_2 \mid x = x_{j_*}, \underline{\beta}_2) dy_2$ in Equation 6.1 we have that

$$P(Y_2 \in I_{2j_2} \mid x \in I_{j_*}, \underline{\beta}_2) = \left[1 + \exp(-\pi(u_{2j_2} - x_{j_*}) / \sqrt{3\beta_{22}}) \right]^{-1} \\ - \left[1 + \exp(-\pi(l_{2j_2} - x_{j_*}) / \sqrt{3\beta_{22}}) \right]^{-1}$$

Using LOS as the IV providing identifying information to this GSME model, would require modeling a segmented linear predictor for γ_{31} . The reason for this is that knowing the ideal GA is 36 weeks, the slope of the line below 36 weeks must be negative since infants with a GA far below 36 weeks would require a longer LOS, where LOS would decrease the closer you came to 36 from the left. Further, as GA increases, infants tend to be more sick, requiring a longer LOS for GA's much greater than 36 weeks, thus requiring a positive slope for values above 36 weeks. The segmented linear predictor with a inflection point at 36 weeks would be given as $\eta_3 = \beta_{30} + (x - 36)\beta_{31}I^{(x < 36)} + (x - 36)\beta_{32}I^{(x \geq 36)}$, where $I^{(x < 36)}$ is an indicator equaling one if $x < 36$ weeks, and is zero otherwise. The second indicator is defined in a similar manner. So assuming $Y_3 \sim \log - \logistic(\lambda_3, p_3)$, the mappings required for Y_3 would be

$$\Leftrightarrow \begin{bmatrix} \gamma_{31} \\ \gamma_{32} \end{bmatrix} = \begin{bmatrix} h_{31}(x; \underline{\beta}_{31}) \\ h_{32}(x; \underline{\beta}_{32}) \end{bmatrix} = \begin{bmatrix} \eta_3 \\ \beta_{33} \end{bmatrix} \\ \begin{bmatrix} \gamma_{31} \\ \gamma_{32} \end{bmatrix} = \begin{bmatrix} T_{31}(\theta_3) \\ T_{32}(\theta_3) \end{bmatrix} = \begin{bmatrix} 1/\lambda_3 \\ 1/p_3 \end{bmatrix}.$$

So for the remaining piece in Equation 6.1, $\int_{I_{3j_3}} f_3(y_3 \mid x = x_{j_*}, \underline{\beta}_3) dy_3$, we would have

$$P(Y_3 \in I_{3j_3} \mid x \in I_{j_*}; \underline{\beta}_3) = \left[1 + \left(\frac{l_{3j_3}}{\eta_3} \right)^{1/\beta_{33}} \right]^{-1} - \left[1 + \left(\frac{u_{3j_3}}{\eta_3} \right)^{1/\beta_{33}} \right]^{-1}.$$

For this example, under the current assumptions, we have as our model parameters, $\underline{\beta} = (\beta_{10}, \beta_{11}, \beta_{22}, \beta_{30}, \beta_{31}, \beta_{32}, \beta_{33}, a, b)'$. At this point, intervals may be chosen for the continuous variables, and once derivatives with respect to the elements of $\underline{\beta}$ are taken for use in the EGNLS procedure, as discussed in Chapter 4, the example is ready to be applied to data, which is available from MCHERDC at the University of Florida. The actual application to data will be done at a future time.

6.2 Application to Alzheimer's Disease

One of the examples discussed in the introductory chapter to this dissertation is the measurement error problem which occurs when one might possibly attempt to determine the association between the likelihood of developing Alzheimer's disease (AD) and the level of aluminum deposits which have accumulated in the brain over time (see, for example, [12]). Recall that a measurement error is created due to the fact that determining a perfectly accurate level of aluminum deposits in the brain is not possible. Assuming one has, for example, some clinical estimate of the amount of aluminum in the brain then this problem could be handled by a GSME model, in the presence of an IV.

The outcome variable of interest, $Y_{r\cdot} = Y_1$, say, would be dichotomous and, obviously, take on the two possible values of AD or no AD. This outcome could be modeled using logistic regression under a generalized linear model. The clinical estimate of the amount of aluminum deposits in the brain would be represented by Y_2 in the GSME model, as the observed value which is contaminated with measurement error. The unattainable, true value for the amount of aluminum deposits would of course be x .

A major difficulty posed in this application is finding a variable which satisfies the requirements of an IV. This would be the third manifest variable, Y_3 , in the GSME model which is the source of identifying information. Because it is not clearly known how aluminum enters the body and accumulates over time, finding

a variable related to the level of aluminum in the brain, or body in general, but independent of both the measurement error in the aluminum level estimate and of the model error in Y_1 (i.e., conditionally independent of Y_1) proves to be a difficult task. Following are a few ideas, but they may rather serve as possible starting places for such future research, as few provide a plausible solution.

The immediate choice for Y_3 would be a second, independent method of estimating aluminum deposits in the brain. As most estimates of aluminum deposits come post-mortem [3], it may be difficult in itself finding one method reliable enough to use as the measured value. A second possibility for Y_3 is the indicator of another disease known to be related to aluminum in the brain. One such example is a disease known as encephalopathy [3], which is a neurological disorder caused by the deterioration of brain functions. Although this deterioration of brain function may not directly be a cause of AD, it is possible that such a disease has some level of correlation with AD. There would be a need for clear, quantifiable indicators of this neurological illness, separate from indications of AD. If encephalopathy is an illness not related to Alzheimer's, it could serve as Y_3 , as it would further be related to the true x , but also independent of the measurement error of x .

The following two variables could serve as IV's, if they proved to provide attainable values in humans. The reason arriving at such values is in question, is that the following studies have been performed on laboratory animals only. They are worth mentioning to serve as starting points for relevant future research. Bushinsky et al. [8] found a relation between the effects of aluminum in the body on bone surface ion composition in laboratory mice. Ion considerations on the bone surface and eroded subsurface were ^{23}Na , ^{27}Al , ^{39}K , and ^{40}Ca . The content of bone surface would be a variable related to the true amount of aluminum in the body, but presumably not related to the model error in the likelihood of developing

AD nor the error in Y_2 . Yokel and O'Callaghan [67] found that an increase in acidic protein in glial fibers in laboratory rabbits was related to aluminum levels injected over time. Glial fibers are contained within the soft tissue which surrounds the central nervous system. Again, this offers a plausible starting place for future research as levels of such acidic protein are related to amounts of aluminum deposits, but may be uncorrelated with the likelihood of developing AD and also uncorrelated with the measurement error in x .

A possible source of aluminum in the body may be due to environmental factors, such as through drinking water. A purely speculative variable for Y_3 could be the location of an individual. If it is determined that different living locations, perhaps between countries, are correlated with different levels of aluminum deposits in the brain, then this variable may serve as a possible IV. The location of living is most likely not correlated with AD. This variable offers possibly the best choice for Y_3 , however, no research has been found stating such a relation.

In medical research, controlling for error free covariates is often important. These variables related to the outcome of interest, have been mentioned throughout this dissertation and denoted by Z . Including error free covariates in the GSME model is discussed as possible future work in Chapter 7. In this application, there are obvious Z covariates which are at least worth mentioning here. These are variables which are not contaminated by measurement error, but are related to the likelihood of developing AD. One such covariate is an individual's sex. It has been found that women are three times as likely to develop AD than are men [27]. Possible evidence of the use of age as a Z covariate is that, at least in women, it has been seen that an increased age at which menopause begins decreases the risk of developing Alzheimer's [25]. Another possible error free covariate that would, ideally, need to be controlled for is race. It is known that there is a decreased risk of Alzheimer's disease associated with having a race of white [59].

This application of the GSME model proves to be difficult, not only due to the lack of a viable IV, but due also to the problems in attempting to estimate aluminum levels in the first place. Alfrey et al. [3] discuss in some detail the difficulty of estimating levels of aluminum in the brain, as well as in muscle tissue and bone. They admit that “a major problem in determining tissue [including brain] aluminum content has been the lack of a sensitive method for aluminum analysis.” [3, p. 184] They used a flameless atomic absorption technique which had a detection limit of 6 ppb for aluminum which is performed at autopsy. It is unknown to the author of this dissertation what improvements have been made in determining aluminum levels or if newer, entirely different methods for estimating aluminum levels in live patients exists. One can clearly see how this important, real world application requires the need of measurement error techniques, which could be handled through the use of a GSME model, to acquire a model for estimation or prediction purposes who’s parameter estimates are not biased, or at least where the bias is reduced.

CHAPTER 7

DISCUSSION AND FUTURE WORK

In closing this dissertation, we end with a chapter briefly summarizing and discussing some of the more crucial aspects of the GSME model defined in Chapter 3, the methodology established in Chapter 4 to estimate the model parameters for the model, and the model's numerous and varied applications. We also briefly discuss plans for work to be completed in the future on this topic as well as a potential alternative solution for the GSME model which may be worth pursuing. Some of these items have only briefly been discussed as future work, others were worked out in detail prior to abandoning them until a future time. For the items where such details have been worked out, an attempt was made to include as many of those details as possible here.

7.1 Discussion of the GSME Model

The GSME model utilizes a realistic means of model identification and is generally applicable. We have seen how many well studied ME models fit into the GSME model framework and its minimum of assumptions makes it an accessible and practical method for real world applications.

The possible impacts of this newly defined, measurement error model are unknown until more research is done in this area. GSME models encompass linear and non-linear models alike, as well as have the ability to handle multinomial-categorical, discrete, and continuous variables. This ME model may be applied to GLM's, but it also improves upon them in the sense that the GSME model is not restricted to have a linear predictor.

The theorems which provide conditions under which the GSME model is identified are significant contributions to the literature on ME models. For

models which tend to be more complex than a simple linear model with normality assumptions, there are no theorems which provide generally applicable rules for model identification. The theorems we provide in Section 3.3 of Chapter 3 may not be generally applicable in all situations, due to the fact that often the joint marginal distribution of the manifest variables is unknown, but at a minimum, they provide a basis for thought. In an early paper, Reiersol [46] provides theorems for identifiability in linear ME models. We now summarize one of his theorems. Assuming the model error and measurement error are distributed as a bivariate normal random vector in a linear model, the intercept, β_1 , is identified if and only if at least one of either x or y , in the equation written without model error, i.e., $y = \beta_0 + x\beta_1$, is not normally distributed [46, p. 380]. It was seen in Section 2.4 of Chapter 2, that the question of identifiability has been addressed only in certain situations, but to the authors knowledge, at the time this dissertation was written, no general theorems exist for large classes of models which are more complex than a linear model.

One of the key assumptions used in the established methodology is that of conditional (given x) independence of the manifest variables. In practice, this assumption may often be questionable, or it simply may be desired to use variables which are known to be conditionally dependent. In the numerous applications of the GSME model, lies a solution to this problem of conditionally dependent manifest variables. The solution has been alluded to throughout Chapters 3 and 4. It was shown in Section 3.2 of Chapter 3 how a multivariate regression model fits into the framework of, and therefore its parameters may be estimated using, a GSME model. By grouping conditionally dependent variables together to form a multivariate manifest vector which is conditionally independent of the remaining manifest variables, we have a solution to the case where the crucial assumption of conditional independence is violated between certain variables.

Although many of the applications exemplified in this dissertation are in the medical field, as mentioned in Chapter 1, measurement error models, and thus the GSME model, can be applied to a number of areas outside of the medical field as well. We saw in Section 3.2 of Chapter 3 how many well studied models fit into the GSME model set-up. These models, obviously, can be found in fields other than the medical/health areas. Such models can be used to describe situations in the social sciences, business applications, agricultural applications, as well as many others. Thus, the GSME model provides a solution to numerous situations where there is an independent variable that is not observed directly, but rather, we only have access to a realization of that variable which is measured with error.

7.2 Monte Carlo Study and Asymptotic Efficiency

An important area of future work is to perform a Monte Carlo study to determine the loss of efficiency in the estimator derived for the GSME model. There is some initial concern that our estimator may lose efficiency due to the fact that information is lost in the categorization step in defining the model as a Latent Class model to apply to EGNLS. We have two potential ways for conducting a Monte Carlo study. One is to assume that all variables are normally distributed, since we can solve the problem in the normal case using methods described in Chapter 2, and compare the MSE from this known solution to the MSE from our method, solving the all normal case through the methodology established for the GSME model in Chapters 3 and 4. The second possible idea is to start with coarse categorizations of the continuous and discrete variables and compare the MSE's as the categorizations become finer and finer, approaching continuous distributions.

Knowing that efficiency is lost due to categorization begs the question, if efficiency can in anyway be regained. We believe that an alternate way of categorizing the non-categorical manifest variables may result in an estimator which is asymptotically efficient, or at least, more efficient than the estimator

under the categorization discussed in Chapter 4. One possibility may be to have the newly defined categories be such that as the sample size increases, the length of the intervals defining the categories decreases slowly enough as to not lead to inconsistent cell proportions, $\hat{\pi}$, but still gain asymptotic efficiency. In this scenario, as n increases, the categorization becomes finer and finer, and thus the newly defined categorical distribution theoretically approaches that of a continuous distribution. We know from Shah [52] that in the diagnostic testing situation, when the binary variables are left unchanged, the resulting estimators from the EGNLS method are asymptotically efficient. For the GSME model under this hypothetical new categorization, knowing that the initial categorical variables remain unchanged, and if the non-categorical variables are categorized in such a way as to approach a continuous distribution with an increase in sample size, the estimators derived from the EGNLS procedure may be more efficient than the estimators derived in Chapter 4, or perhaps, even asymptotically fully efficient.

7.3 The GSME Model with Error-Free Covariates

It was seen in Chapter 2, that many of the existing solutions to measurement error problems have the capacity to include covariates which are free of measurement error. Recall that these covariates have previously been denoted by $\underline{Z}$. As was previously mentioned, this would be an important extension of the GSME model, as there are many applications where controlling for such variables would be required. For example, in the RERF application, there is information contained in the error free covariates age at time of bomb, sex, and city. These are variables which are not related to the amount of radiation exposure, but they are related to the outcome of interest, time until the diagnosis of leukemia. A valid analysis should control for these variables when estimating the effect of radiation dose. The potential problem with this extension is that in applications, a sample large enough

to allow estimation of the π parameters using the method described in Chapter 4 may not be available.

The basic idea behind including error free covariates to the GSME model is not complex. To allow full generality continuous, discrete, and multinomial-categorical error free variables would be allowed. The non-categorical $\underline{Z}$ covariates would be categorized in a similar manner to that described for the manifest variables in Chapter 4. In the presence of these variables, we would assume that all conditional, given x and $\underline{Z}$, distributions of the manifest variables would be known. Further, the distribution of x given $\underline{Z}$ would also be known.

The joint marginal distribution of the manifest variables would initially be written in a similar form to Equation 4.1, except the right hand side would be multiplied by the distributions of the $\underline{Z}$ variates and also would be integrated over the ranges of all continuous error free covariates and summed over the discrete values or categories of all discrete and multinomial categorical error free variables, respectively. The same complications arise when considering the analog equation to Equation 4.1 in the presence of $\underline{Z}$ variables, in that the joint marginal distribution all of the $\underline{Y}$'s would in most cases still be unknown and the integration required on the right hand side of the analog equation would still most likely not exist in a closed form. There is a change in the parameter vector of interest in that it would also now contain the variables which describe the relation between the $\underline{Y}$'s and the $\underline{Z}$'s, x and the $\underline{Z}$'s, and the distributional parameters of the marginal distributions of the error free covariates.

Since the theorems on identifiability, presented in Section 3.3 of Chapter 3, are applied to the joint distribution of the manifest variables, these results would still hold in the presence of error free covariates, because such variables are not present in the joint distribution of all the $\underline{Y}$'s. One difference, as mentioned above, is that the vector of model parameters, $\underline{\beta}$, would be of greater dimension. This may,

perhaps, make it more difficult for a model in the presence of error free covariates to be identified.

For each possible cross-classification of the categories of the categorized $\underline{Z}$ covariates, we would obtain the multinomial relation $\hat{\pi} = \pi(\underline{\beta}) + \underline{\varepsilon}$. Therefore, for fixed levels of the error free covariates, we could estimate the model parameters. The possible set back is that there likely would not be available enough observations to fill all possible cells in the multinomial relation $\hat{\pi} = \pi(\underline{\beta}) + \underline{\varepsilon}$ for every possible combination of the categorized error free covariates. All combinations of cross-classifications of the $\underline{Z}$ covariates together form this nonlinear system of equations which is solved by the EGNLS procedure. Assuming a sample size exists that is large enough to fill all possible cells in the entire vector of multinomial proportions, the estimation procedure remains the same, and the asymptotic properties established at the end of Chapter 4 would still hold for the parameter estimates in the presence of $\underline{Z}$ covariates. The asymptotic theorems would generalize to this case easily since those theorems are based on the vector of model parameters, $\underline{\beta}$, and their relation to the nonlinear system of multinomial proportions. Those relations are not directly effected by the inclusion of error free covariates.

Since sample sizes in applications would most likely not be large enough to allow for parameter estimation, future work in this area would most likely be simulation-based or only applied to examples which require a minimum number of categories for the variables. This second possible area for researching the inclusion of error free covariates would most likely best be applied to variables which are already categorical, having few categories. Since categorization would not be required, the large amount of information that would be lost due to the categorization of variables into a small number of categories, would be retained.

7.4 An Alternate Method

Possible future work could also be the further development of a second, general IV estimation method for GSME models. This potential second method differs from the one presented in this dissertation because it begins with an approximation that may lead to a slight bias in parameter estimation, however, there is no potential for a loss of efficiency. We had begun work on this second method but realized early that from this potential method alone, one could not arrive at consistent initial estimates for use in the EGNLS procedure. This is due to the approximation that is made in the initial step of this method. It was determined that consistent estimates would need to be obtained from another method, e.g., the method described at the end of Chapter 4, and used as the initial estimates under this alternate method. Since this method could not be used independently of any other, it was decided that the method described in this dissertation was of greater possible value, because it alone could be used to consistently estimate the model parameters. The idea behind the alternative method and the details which had been worked out earlier are given below.

This method will be able to handle continuous, discrete, and multinomial-categorical manifest variables, just as the method presented in this dissertation. Each manifest variable is assumed to have a known conditional (given x) distribution, generically denoted as $f_r(\underline{y}_r \mid x; \underline{\theta}_r)$. The models for these variables would be specified through the same mappings as those given in Chapter 3. Recall that these mappings, for each manifest variable may be written together in vector form as $\underline{\theta}_r = \underline{T}_r^{-1} \left(\underline{h}_r(x; \underline{\beta}_r) \right)$ as seen in Equation 3.4. This relation allows us to substitute x and the model parameters, $\underline{\beta}_r$, for each r , which describe the relation between $\underline{Y}_r$ and x , in place of the conditional distribution parameters, $\underline{\theta}_r$. These mappings have the same properties as those defined in Chapter 3.

The distribution of x has a range of possible values from negative infinity to positive infinity. In the case for our motivating example, where x is the true, unknown radiation dose, the range of x is from zero to infinity. The first step is to apply the definition of the Riemann integral (see, for example, [34, p. 206]) to rewrite the right hand side of Equation 4.1, which describes the joint distribution of the manifest variables prior to categorization. As in the method presented above, we initially state the joint marginal distribution in its full generality, but for notational convenience, we show the details assuming univariate outcomes where all variables satisfy the conditional independence assumption. Recall that the joint distribution is

$$f(\underline{y}_1, \underline{y}_2, \dots, \underline{y}_p) = \int_x \prod_{r=1}^p f_r(\underline{y}_r | x, \underline{\theta}_r) f(x | \underline{\theta}_x) dx,$$

where $f_r(\underline{y}_r | x, \underline{\theta}_r)$ may be the conditional distribution function for a continuous, discrete, or multinomial categorical manifest variable or any combination of dependent variables which are grouped together to form a conditionally independent group. First we must write that integral as the limit of an integral over a bounded region as follows:

$$\int_{-\infty}^{\infty} \prod_{r=1}^p f_r(\underline{y}_r | x, \underline{\theta}_r) f(x | \underline{\theta}_x) dx = \lim_{b \rightarrow \infty} \int_{-b}^b \prod_{r=1}^p f_r(\underline{y}_r | x, \underline{\theta}_r) f(x | \underline{\theta}_x) dx. \quad (7.1)$$

Here, $f(x | \underline{\theta}_x)$ represents the extended version of the density defined as the density function of x over its domain and zero elsewhere. Now let the interval $[-b, b]$ be partitioned into a finite number of subintervals with endpoints defined by the set of points $P = \{x_0, x_1, \dots, x_k\}$ such that $-b = x_0 < x_1 < \dots < x_k = b$. Further, let the length of each interval be $\frac{b-(-b)}{k}$ so that the set P contains the $k+1$ numbers $x_j = -b + j \frac{b-(-b)}{k}$, $j = 0, 1, \dots, k$. Define $h_j = x_j - x_{j-1}$, $j = 1, 2, \dots, k$, and m_j as the midpoint in the subinterval $[x_{j-1}, x_j]$, $j = 1, 2, \dots, k$. We define different notation and a different procedure for the categorization of x from that described

in Chapter 4 because this potential solution utilizes the definition of the Riemann integral and this new notation is needed to do so, since the length of the intervals here must approach zero as $n \rightarrow \infty$. Note, also, that the intervals for x in this case are required to be of equal length, whereas the categorization previously described for x did not require intervals of equal length. The definition of the Riemann integral applied to the right hand side of Equation 7.1 gives

$$\lim_{b \rightarrow \infty} \int_{-b}^b \prod_{r=1}^p f_r(\underline{y}_r \mid x, \underline{\theta}_r) f(x \mid \underline{\theta}_x) dx = \lim_{b \rightarrow \infty} \lim_{k \rightarrow \infty} \sum_{j=1}^k \prod_{r=1}^p f_r(\underline{y}_r \mid x = m_j, \underline{\theta}_r) f(m_j \mid \underline{\theta}_x) h_j. \quad (7.2)$$

We will return to Equation 7.2, but first we shall manipulate Equation 4.1, the joint marginal distribution of the manifest variables, making use of the multinomial distribution through categorized versions of the continuous and discrete manifest variables, and those variables which were already categorical.

For notational convenience, we shall assume all manifest variables are univariate and conditionally independent given x . Without loss of generality, order the manifest variables such that Y_r , $r = 1, 2, \dots, p'$, are the continuous variables, Y_r , $r = p' + 1, p' + 2, \dots, p''$, are the “other” discrete variables, and Y_r , $r = p'' + 1, p'' + 2, \dots, p$, are the categorical variables which initially follow a multinomial distribution. First, categorize each continuous and “other” discrete manifest variable. Recall that the “other” discrete variables are those discrete variables which do require further categorization, i.e., do not initially follow a multinomial distribution. Categorize each Y_r , $r = 1, 2, \dots, p''$, into C_r categories, in a similar manner as described in Chapter 4 for the GSME model. This categorization results in univariate intervals, I_{rj_r} , having endpoints (l_{rj_r}, u_{rj_r}) , $r = 1, 2, \dots, p''$, $j_r = 1, 2, \dots, C_r$, however, now denote length(I_{rj_r}) = h_{rj_r} . This different notation is due to the fact that the lengths of the intervals for this method have slightly different properties than those for the method given in this dissertation. Previously,

the interval lengths were allowed to remain fixed as $n \rightarrow \infty$, in this case, the intervals must shrink to a length of zero as $n \rightarrow \infty$. So here, C_r will depend on n . The reason for this will become apparent in details which follow. As before, the categorical manifest variables that follow a multinomial distribution, i.e., Y_r , $r = p'' + 1, p'' + 2, \dots, p$, will remain unchanged and the number of their categories will again be denoted by C_r , $r = p'' + 1, p'' + 2, \dots, p$, where the categories are indexed by j_r , $j_r = 1, 2, \dots, C_r$.

Further, as in Chapter 4, let

$$\pi_{j_1 j_2 \dots j_p} = P(Y_1 \in I_{1j_1}, \dots, Y_{p''} \in I_{p''j_{p''}}, Y_{p''+1} = j_{p''+1}, \dots, Y_p = j_p).$$

Define $\underline{\pi}$ to be the vector of $N = \prod_{r=1}^p C_r$ distinct categories in the cross-classification of the categorized versions of continuous and discrete Y_r , $r = 1, 2, \dots, p''$, and the categorical Y_r , $r = p'' + 1, p'' + 2, \dots, p$. That is, $\underline{\pi}$ is the vector of distinct multinomial probabilities of falling into the categories defined by the cross-classifications' cells. Assume $h_{r,j_r} = o(1/n^\delta)$ and that the first endpoint and last endpoint for each Y_r , $r = 1, 2, \dots, p''$, are such that $l_{r1} = o(-n^\delta)$ and $u_{rC_r} = o(n^\delta)$, respectively, where $\delta > 0$ is sufficiently small so that $n_{j_1 j_2 \dots j_p} \rightarrow \infty$ as $n \rightarrow \infty$, for all $j_1, j_2, \dots, j_p$. Then, it can be shown that

$$f(y_{1m_{j_1}}, y_{2m_{j_2}}, \dots, y_{p''m_{j_{p''}}}, y_{p''+1}, \dots, y_p) = \lim_{\max_{j_1} (h_{1j_1}) \rightarrow 0} \dots \lim_{\max_{j_{p''}} (h_{p''j_{p''}}) \rightarrow 0} \frac{\pi}{\prod_{r=1}^{p''} h_{rj_r}},$$

where $y_{rm_{j_r}}$, $r = 1, 2, \dots, p''$, is chosen to be the midpoint of the j_r th interval of the continuous and discrete Y_r , $r = 1, 2, \dots, p''$, and $\underline{\pi}$, again, is the vector of distinct probabilities of the multinomial distribution defined by the cross-classification of the categorized continuous and discrete manifest variables and those which were already categorical. Therefore, let

$$\underline{y} = (y_{1m_{j_1}}, y_{2m_{j_2}}, \dots, y_{p''m_{j_{p''}}}, y_{p''+1}, \dots, y_p)'$$

and we define $\widehat{f}(\underline{y}) = \widehat{\pi}(1/\prod_{r=1}^{p''} h_{r,j_r})$, where the sample proportions, $\widehat{\pi}$, are unbiased and consistent estimates of the cell probabilities, $\underline{\pi}$. Add and subtract $\widehat{f}(\underline{y})$ to the left hand side of Equation 4.1 giving

$$\begin{aligned} & \left[f_Y(y_1, y_2, \dots, y_p) - \widehat{f}(\underline{y}) \right] + \widehat{f}(\underline{y}) = \\ & \int_x \prod_{r=1}^{p'} f_r(y_r | x, \underline{\theta}_r) \prod_{r=p'+1}^{p''} p_r(y_r | x, \underline{\theta}_r) \prod_{r=p''+1}^p P(Y_r = j_r | x, \underline{\theta}_r) f(x | \underline{\theta}_x) dx, \end{aligned}$$

where $f_r(y_r | x, \underline{\theta}_r)$ is a probability density function for a continuous Y_r , $p_r(y_r | x, \underline{\theta}_r)$ is a probability mass function for a discrete Y_r , and $P(Y_r = j_r | x, \underline{\theta}_r)$ represents conditional multinomial cell probabilities for a categorical Y_r . This implies

$$\begin{aligned} & \widehat{f}(\underline{y}) = \\ & \int_x \prod_{r=1}^{p'} f_r(y_r | x, \underline{\theta}_r) \prod_{r=p'+1}^{p''} p_r(y_r | x, \underline{\theta}_r) \prod_{r=p''+1}^p P(Y_r = j_r | x, \underline{\theta}_r) f(x | \underline{\theta}_x) dx \\ & + \left[\widehat{f}(\underline{y}) - f_Y(y_1, y_2, \dots, y_p) \right], \end{aligned} \quad (7.3)$$

where $\widehat{f}(\underline{y}) - f_Y(y_1, y_2, \dots, y_p) \doteq (1/\prod_{r=1}^{p''} h_{r,j_r})\widehat{\pi} - f_Y(y_1, y_2, \dots, y_p)$ and $\widehat{\pi} = \underline{\pi} + \underline{\varepsilon}$ where $\underline{\varepsilon} = O_p(1/\sqrt{n})$. So, from Equation 7.3, we have

$$\begin{aligned} \widehat{f}(\underline{y}) &= (1/\prod_{r=1}^{p''} h_{r,j_r})\widehat{\pi} \\ &= \int_x \prod_{r=1}^{p'} f_r(y_r | x, \underline{\theta}_r) \prod_{r=p'+1}^{p''} p_r(y_r | x, \underline{\theta}_r) \prod_{r=p''+1}^p P(Y_r = j_r | x, \underline{\theta}_r) f(x | \underline{\theta}_x) dx \\ &\quad + (1/\prod_{r=1}^{p''} h_{r,j_r})(\underline{\pi} + \underline{\varepsilon}) - f_Y(y_1, y_2, \dots, y_p). \end{aligned}$$

Multiply both sides by $\prod_{r=1}^{p''} h_{r,j_r}$ giving

$$\begin{aligned} \widehat{\pi} &= \int_x \prod_{r=1}^{p'} f_r(y_r | x, \underline{\theta}_r) \prod_{r=p'+1}^{p''} p_r(y_r | x, \underline{\theta}_r) \prod_{r=p''+1}^p P(Y_r = j_r | x, \underline{\theta}_r) f(x | \underline{\theta}_x) \prod_{r=1}^{p''} h_{r,j_r} dx \\ &\quad + [\underline{\pi} - f_Y(y_1, y_2, \dots, y_p) \prod_{r=1}^{p''} h_{r,j_r}] + \underline{\varepsilon}, \end{aligned} \quad (7.4)$$

where, again,

$$\underline{\varepsilon} = \widehat{\underline{\pi}} - \underline{\pi} = O_p(1/\sqrt{n})$$

and

$$\underline{\pi} - f_Y(y_1, y_2, \dots, y_p) \prod_{r=1}^{p''} h_{r_{j_r}} = O\left(\left(\prod_{r=1}^{p''} h_{r_{j_r}}\right)^{\delta+1}\right)$$

for some $\delta > 0$. This second fact can easily be seen in the univariate case, for a single variable, where $\lim_{h_1 \rightarrow 0} (\frac{\pi}{h_1} - f(y_1)) = 0$ implies there exists a $\delta > 0$ such that $(\frac{\pi}{h_1} - f(y_1)) = O(h_1^\delta)$ implying $(\pi - f(y_1)h_1) = O(h_1^\delta)h_1 = O(h_1^{\delta+1})$, where recall that h_1 is the length of an interval. Finally, plugging the right hand side of Equation 7.2 into the right hand side of Equation 7.4 gives

$$\begin{aligned} \widehat{\underline{\pi}} &= \lim_{b \rightarrow \infty} \lim_{k \rightarrow \infty} \sum_{j=1}^k \prod_{r=1}^{p'} f_r(y_r \mid x = m_j, \underline{\theta}_r) \\ &\quad \times \prod_{r=p'+1}^{p''} p_r(y_r \mid x = m_j, \underline{\theta}_r) \prod_{r=p''+1}^p P(Y_r = j_r \mid x = m_j, \underline{\theta}_r) f(m_j \mid \underline{\theta}_x) h_j \prod_{r=1}^{p''} h_{r_{j_r}} \\ &\quad + O\left(\left(\prod_{r=1}^{p''} h_{r_{j_r}}\right)^{\delta+1}\right) + \underline{\varepsilon}. \end{aligned} \quad (7.5)$$

Dropping the limits on the right hand side of the preceding equation gives

$$\begin{aligned} \widehat{\underline{\pi}} &= \sum_{j=1}^k \prod_{r=1}^{p'} f_r(y_r \mid x = m_j, \underline{\theta}_r) \\ &\quad \times \prod_{r=p'+1}^{p''} p_r(y_r \mid x = m_j, \underline{\theta}_r) \prod_{r=p''+1}^p P(Y_r = j_r \mid x = m_j, \underline{\theta}_r) f(m_j \mid \underline{\theta}_x) h_j \prod_{r=1}^{p''} h_{r_{j_r}} \\ &\quad + O\left(\left(\prod_{r=1}^{p''} h_{r_{j_r}}\right)^{\delta+1}\right) \\ &\quad + \underline{\varepsilon} + o\left(\prod_{j=1}^k h_j\right) + B_b, \end{aligned} \quad (7.6)$$

where we may assume $O\left(\left(\prod_{r=1}^{p''} h_{r_{j_r}}\right)^{\delta+1}\right)$ is negligible, for sufficiently small $h_{r_{j_r}}$, $r = 1, 2, \dots, p''$, $j_r = 1, 2, \dots, C_r$, $o(\prod_{j=1}^k h_j)$ is negligible, for sufficiently small h_j , $j = 1, 2, \dots, k$, and B_b represents the integral over the range of negative infinity

to negative b and b to infinity, i.e., the sum of the tail areas, which is negligible for sufficiently large b . This is due to the fact that $\prod_r f_r(y_r | x, \underline{\theta}_r) \prod_r p_r(y_r | x, \underline{\theta}_r) \prod_r P(Y_r = j_r | x, \underline{\theta}_r) f(x | \underline{\theta}_x)$ is the joint distribution of $\underline{y}$, $r = 1, 2, \dots, p$, and x and therefore is theoretically integrable with respect to the continuous variable x . At this point in the estimation procedure, the model parameters would enter the right hand side of Equation 7.6 by substituting for the conditional distribution parameters, those mappings which are functions of x and the model parameters, i.e., substituting the elements of $\underline{\theta}_r$ with the elements of $T_r^{-1} \underline{h}_r(x; \underline{\beta}_r)$. Then the conditional distributions could be written, for example for a continuous manifest variable, as $f_r(y_r | x, \underline{\beta}_r)$, indicating the substitution has occurred.

Since the distributions of the manifest variables are assumed to be known, specific forms for any example can be inserted into the right hand side of Equation 7.6 in terms of the model parameters which enter nonlinearly from these distributions. The resulting system of equations is solved via iterative generalized nonlinear least squares using the same method as that described in Chapter 4, by ignoring the $O\left((\prod_{r=1}^{p''} h_{r_{j_r}})^{\delta+1}\right)$, $o(\prod_{j=1}^k h_j)$, and B_b terms.

The resulting estimators should be biased. The reason for the bias in the estimators comes from the fact that the $o(\prod_{j=1}^k h_j)$ and B_b terms are dropped in Equation 7.6 causing a slight shift in the likelihood. The $o(\prod_{j=1}^k h_j)$ term and the B_b term are needed for Equation 7.6 to be the correct mean function. The univariate case was presented for notational ease, however, just as in the method presented in this dissertation, the concept can be extended to the multivariate case, allowing for groups of manifest variables conditionally dependent with each other, but as a group, conditionally independent, given x , of the remaining variables.

REFERENCES

- [1] Agresti, A. (1990) *Categorical Data Analysis*, New York: Wiley.
- [2] Aitkin, M. and Rocci, R. (2002), "A General Maximum Likelihood Analysis of Measurement Error in Generalized Linear Models," *Statistics and Computing*, **12**(2), 163-174.
- [3] Alfrey, A. C., LeGendre, G. R., and Kaehny, W. D. (1976), "The Dialysis Encephalopathy Syndrome: Possible Aluminum Intoxication," *The New England Journal of Medicine*, **294**(4), 184-188.
- [4] Alzheimer's Disease and Related Disorders Association, Inc. (2001), *Alzheimer's Disease Statistics*, Chicago: Author.
- [5] Amemiya, Y. (1985), "Instrumental Variable Estimation for the Nonlinear Errors-in-Variables Model," *Journal of Econometrics*, **28**, 273-289.
- [6] Armstrong, B. (1985), "Measurement Error in the Generalized Linear Model," *Communications in Statistics: Simulation and Computation*, **14**(3), 529-544.
- [7] Burquet, A., Monnet, E., Roth, P., Hirn, F., Vouaillat, C., Lecourt-Ducet, M. et al. (2000), "Neurodevelopmental Outcomes of Premature Infants Born Less than 33 Weeks of Gestational Age and Not Cerebral Palsy at the Age of 5 Years," *Archives de Pediatrie*, **7**(4), 357-368.
- [8] Bushinsky, D. A., Sprague, S. M., Hallegot, P., Girod, C., Chabala, J. M., and Levi-Setti, R. (1995), "Effects of Aluminum on Bone Surface Ion Composition," *Journal of Bone and Mineral Research*, **10**(12), 1988-1997.
- [9] Buzas, J. S. (1997), "Instrumental Variable Estimation in Nonlinear Measurement Error Models," *Communications in Statistics: Theory and Methodology*, **26**(12), 2861-2877.
- [10] Buzas, J. S. and Stefanski, L. A. (1996), "Instrumental Variable Estimation in Generalized Linear Models," *Journal of the American Statistical Association*, **91**(435), 999-1006.
- [11] Buzas, J. S. and Stefanski, L. A. (1996), "Instrumental Variable Estimation in a Probit Measurement Error Model," *Journal of Statistical Planning and Inference*, **55**, 47-62.
- [12] Campbell, A. (2002), "The Potential Role of Aluminum in Alzheimer's Disease," *Nephrology, Dialysis, Transplantation*, **17**, 17-20.

- [13] Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*, London: Chapman and Hall.
- [14] Carroll, R. J., Spiegelman, C. H., Lan, K., Bailey, K. and Abbott, R. (1984), "On Errors in Variables for Binary Regression Models," *Biometrika*, **71**(1), 19-25.
- [15] Carroll, R. J. and Stefanski, L. A. (1990), "Approximate Quasi-Likelihood Estimation in Models with Surrogate Predictors," *Journal of the American Statistical Association*, **85**(411), 652-663.
- [16] Carter, R. L. (1976), *Instrumental Variable Estimation of the Simple Errors in Variables Model*, unpublished Ph. D. dissertation, Iowa State University, Department of Statistics.
- [17] Carter, R. L. (1981), "Restricted Maximum Likelihood Estimation of Bias and Reliability in the Comparison of Several Measuring Methods," *Biometrics*, **37**(4), 733-741.
- [18] Carter, R. L. and Fuller, W. A. (1980), "Instrumental Variable Estimation of the Simple Errors-in-Variables Model," *Journal of the American Statistical Association*, **75**(371), 687-692.
- [19] Cook, J. R. and Stefanski, L. A. (1994), "Simulation-Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, **89**(428), 1314-1328.
- [20] Dennis, J. E. and Schnabel, R. B. (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, New Jersey: Prentice-Hall.
- [21] Foutz, R.V. (1977), "On The Unique Consistent Solutions to the Likelihood Equations," *Journal of the American Statistical Association*, **72**(357), 147-148.
- [22] Fuller, W. A. (1976) *Introduction to Statistical Time Series*, New York: Wiley.
- [23] Fuller, W. A. (1980), "Properties of Some Estimators for the Errors in Variables Model," *The Annals of Statistics*, **8**(2), 407-422.
- [24] Fuller, W. A. (1987) *Measurement Error Models*, New York: Wiley.
- [25] Geerlings, M. I., Ruitenberg, A., Witteman, J. C., Van Swieten, J. C., Hofman, A., Van Duijn, C. M., Breteler, M. M. and Launer, L. J. (2001), "Reproductive Period and Risk of Dementia in Postmenopausal Women," *Journal of the American Medical Association*, **285**, 1475-1481.
- [26] Gleser, L. J. (1990), "Improvements of the Naive Approach to Estimation in Nonlinear Errors-in-Variables Regression Models," In *Contemporary Mathematics Volume 112: Statistical Analysis of Measurement Error Models and Applications*, eds. P. J. Brown and W. A. Fuller, 99-114.

- [27] Greene, R. A. (2000), "Estrogen and Cerebral Blood Flow; A Mechanism to Explain the Impact of Estrogen on the Incidence and Treatment of Alzheimer's Disease," *International Journal of Fertility and Women's Medicine*, 45, 253-257.
- [28] Greenland, S. (2000), "An Introduction to Instrumental Variables for Epidemiologists," *International Journal of Epidemiology*, 29(4), 722-729.
- [29] Griliches, Z. and Ringstad, V. (1970), "Error-in-the-Variables Bias in Nonlinear Contexts," *Econometrica*, 38(2), 368-370.
- [30] Hack, M. and Fanaroff, A. A. (2000), "Outcomes of Children of Extremely Low Birthweight and Gestational Age in the 1990's," *Seminars in Neonatology*, 5(2), 89-106.
- [31] Hartley, H. O. (1961), "The Modified Gauss-Newton Method for the Fitting of Non-linear Regression Functions by Least Squares," *Technometrics*, 3, 269-280.
- [32] Honig, L. S. and Mayeux, R. (2001), "Natural History of Alzheimer's Disease," *Aging-Clinical and Experimental Research*, 13, 171-182.
- [33] Kalbfleisch, J. D. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data*, New York: Wiley.
- [34] Khuri, A. I. (1993) *Advanced Calculus with Applications in Statistics*, New York: Wiley.
- [35] Klein, J. P. and Moeschberger, M. L. (1997) *Survival Analysis Techniques for Censored and Truncated Data*, New York: Springer.
- [36] Klepper, S. and Leamer, E. E. (1984), "Consistent Sets of Estimates for Regressions with Errors in all Variables," *Econometrica*, 52(1), 163-183.
- [37] Larsen, T., Nguyen, T. H., Griesen, G., Engholm, G., and Moller, H. (2000), "Does a Discrepancy Between Gestational Age Determined by Biparietal Diameter and Last Menstrual Period Sometimes Signify Early Intrauterine Growth Retardation?," *BJOG: An International Journal of Obstetrics and Gynaecology*, 107(2), 238-244.
- [38] Lazarsfeld, P. F. and Henry, N. W. (1968) *Latent Structure Analysis*, Boston: Houghton Mifflin.
- [39] Mackanjee, H. R., Iliescu, B. M., Dawson, W. B. (1996), "Assessment of Postnatal Gestational Age Using Sonographic Measurement of Femur Length," *Journal of Ultrasound in Medicine*, 15(2), 115-120.
- [40] McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, Boca Raton: Chapman and Hall/CRC.

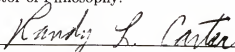
- [41] Neriishi, K., Stram, D., Vaeth, M., Mizuno, S. and Akiba, S. (1991), "The Observed Relationship Between the Occurrence of Acute Radiation Effects and Leukemia Mortality among A-Bomb Survivors," *Radiation Research*, **125**(2), 206-213.
- [42] Pierce, D. A., Stram, D. O., and Vaeth, M. (1990), "Allowing for Random Errors in Radiation Dose Estimates for the Atomic Bomb Survivor Data," *Radiation Research*, **123**(3), 275-284.
- [43] Pierce, D. A., Stram, D. O., Vaeth, M., and Schafer, D. (1992), "The Errors in Variables Problem: Considerations Provided by the Radiation Dose-Response Analyses of the A-Bomb Survivor Data," *Journal of the American Statistical Association*, **87**(418), 351-359.
- [44] Prentice, R. L. (1982), "Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model," *Biometrika*, **69**(2), 331-342.
- [45] Rao, C. R. (1973) *Linear Statistical Inference and its Applications*, New York: Wiley.
- [46] Reiersol, O. (1950), "Identifiability of a Linear Relation Between Variables which are Subject to Error," *Econometrica*, **18**, 375-389.
- [47] Resnick, M. B., Ariet, M., Carter, R. L., Bucciarelli, R. L., Furlough, R., Evans, J. H., et al. (1988), "Prospective Pricing Model for Neonatologists and Obstetricians in Tertiary Care Centers," *Pediatrics*, **82**, 442-446.
- [48] Rudin, W. (1976) *Principles of Mathematical Analysis*, Singapore: McGraw-Hill.
- [49] Schott, J. R. (1997) *Matrix Analysis for Statistics*, New York: Wiley.
- [50] Sen, P. K. and Singer, J. M. (1993) *Large Sample Methods in Statistics: An Introduction with Applications*, New York: Chapman and Hall.
- [51] Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*, New York: Wiley.
- [52] Shah, N. (1998), *Estimated Generalized Nonlinear Least Squares for Latent Class Analysis of Diagnostic Tests*, unpublished Ph. D. dissertation, University of Florida, Department of Statistics.
- [53] Smith, L. N., Dayal, V. H., and Monga, M. (1999), "Prior Knowledge of Obstetric Gestational Age and Possible bias of Ballard Score," *Obstetrics and Gynecology*, **93**(5), 712-714.
- [54] Sposto, R., Stram, D. O., and Awa, A. A. (1991), "An Estimate of the Magnitude of Random Errors in the DS86 Dosimetry from Data on Chromosome Aberrations and Severe Epilation," *Radiation Research*, **128**(2), 157-169.

- [55] Stefanski, L. A. (1985), "The Effects of Measurement Error on Parameter Estimation," *Biometrika*, **72**(3), 583-592.
- [56] Stefanski, L. A. (2000), "Measurement Error Models," *Journal of the American Statistical Association*, **95**(452), 1353-1358.
- [57] Stefanski, L. A. and Boos, D. D. (2002), "The Calculus of M-Estimation," *The American Statistician*, **56**(1), 29-38.
- [58] Stefanski, L. A. and Buzas, J. S. (1995), "Instrumental Variable Estimation in Binary Regression Measurement Error Models," *Journal of the American Statistical Association*, **90**(430), 541-550.
- [59] Tang, M. X., Cross P., Andrews, H., Jacobs, D. M., Small, S., Bell, K., Merchant, C., Lantigua, R., Costa, R., Stern, Y., and Mayeux, R. (2001), "Incidence of AD in African-Americans, Caribbean Hispanics, and Caucasians in Northern Manhattan," *Neurology*, **56**, 49-56.
- [60] Thompson, J. R., Carter, R. L., Ariet, M., Roth, J., Ross, N., and Resnick, M. B. (2001), "An Assessment of Birth Weight and Social Demographic Effects on Early Childhood Disability or Developmental Delay," *Florida Health Care Journal*, **2**(1), 7-15.
- [61] Thompson, J. R., Carter, R. L., Edwards, A. R., Roth, J., Ariet, M., Ross, N., and Resnick, M. B., "A Population Based Study of the Effects of Birth Weight on Early Developmental Delay or Disability in Children," *American Journal of Perinatology*, To appear.
- [62] Thoresen, M. and Laake, P. (1999), "Instrumental Variable Estimation in Logistic Measurement Error Models by Means of Factor Scores," *Communications in Statistics: Theory and Methodology*, **28**(2), 297-313.
- [63] Upton, G. and Cook, I. (2002) *A Dictionary of Statistics*, New York: Oxford University Press.
- [64] Wolfe, M. A. (1978) *Numerical Methods for Unconstrained Optimization: an Introduction*, New York: Van Nostrand Reinhold.
- [65] Wolter, K. M. and Fuller, W. A. (1982), "Estimation of Nonlinear Errors in Variables Models," *The Annals of Statistics*, **10**(2), 539-548.
- [66] Woods, N. S., Marlow, N., Costeloe, K., Gibson, A. T., and Wilkinson, A. R. (2000), "Neurologic and Developmental Disabilities After Extremely Preterm Birth," *New England Journal of Medicine*, **343**(6), 378-384.
- [67] Yokel, R. A. and O'Callaghan, J. P. (1998), "An Aluminum-Induced Increase in GFAP is Attenuated by Some Chelators," *Neurotoxicology and Teratology*, **20**(1), 55-60.

BIOGRAPHICAL SKETCH

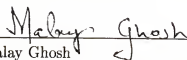
Jeffrey Ray Thompson was born June 17, 1973, to Ron and Ellie Thompson in Natrona Heights, Pennsylvania. He has one elder sister, Jennifer. After living in both Pennsylvania and Ohio for a few short years, Jeff's family relocated to Orlando, Florida. Jeff spent the majority of his adolescence growing up in Orlando, where he graduated from Lake Brantley High School in 1991. He went to Stetson University in Deland, Florida, where he obtained a Bachelor of Science degree in mathematics in 1995. It was while attending Stetson that Jeff developed his interest for the field of statistics. His Master of Science degree in statistical computing was granted by the University of Central Florida in Orlando in 1997 and it was through the opportunity to teach undergraduate statistics while working on his master's degree that Jeff developed his passion for teaching. Desiring a career in academia, Jeff decided to pursue his Ph.D. in statistics at the University of Florida, in Gainesville, where he intends to graduate in 2003. He has accepted an Assistant Professor teaching position with the Department of Statistics at North Carolina State University, in Raleigh, NC.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



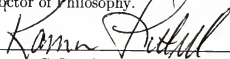
Randy L. Carter, Chair
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



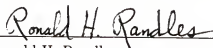
Malay Ghosh
Distinguished Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



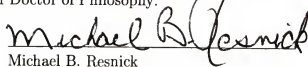
Ramon C. Littell
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Ronald H. Randles
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Michael B. Resnick
Professor of Pediatrics

This dissertation was submitted to the Graduate Faculty of the Department of Statistics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 2003

Dean, Graduate School